7th European Symposium on
Computational Intelligence and Mathematics

ESCIM 2015

Cádiz, Spain

October 7th-10th, 2015

Proceedings

Editors:
László Kóczy, Jesús Medina

Associate Editors:
María Eugenia Cornejo-Piñero, Juan Carlos Díaz-Moreno
Janusz Kacprzyk, Valentín Liñeiro-Barea
Eloísa Ramírez-Poussa, María José Benítez-Caballero

Proceedings of ESCIM 2015

Edition: 1st
First published: 2015

ISBN: 978-84-608-2823-5

Published and printed by:
Universidad de Cádiz (Dept. Matemáticas), Spain

Organization

General Chairs

László T. Kóczy	Univ. Széchenyi István, Gÿor, Hungary
Janusz Kacprzyk	Polish Academy of Sciences, Warsaw, Poland
Jesús Medina	Universidad de Cádiz, Spain

International Program Commitee

Pedro Cabalar	Universidad de A Coruña, Spain
Agata Ciabattoni	TU Wien, Austria
Davide Ciucci	University of Milano-Bicocca, Italy
Bernard De Baets	University of Gante, Belgium
Chris Cornelis	Universidad de Granada, Spain
Christian G. Fermueller	TU Wien, Austria
Péter Foldesi	Univ. Széchenyi István, Hungary
Lluis Godo	Artificial Intelligence Research Institute, Spain
László T. Kóczy	Univ. Széchenyi István, Hungary
Stanislav Krajci	UPJS Kosice, Slovakia
Ondrej Kridlo	UPJS Kosice, Slovakia
Piotr Kulczycki	Cracow University of Technology, Poland
Inmaculada Medina	Universidad de Cádiz, Spain
Jesús Medina	Universidad de Cádiz, Spain
Manuel Ojeda-Aciego	Universidad de Málaga, Spain
David Pearce	Universidad Politécnica de Madrid, Spain
Jozef Pócs	Slovak Academy of Sciences, Slovakia
Claudiu Pozna	Transilvania University of Brasov, Romania
Alex Tormási	Univ. Széchenyi István, Hungary
Esko Turunen	Tampere University of Technology, Finland
Agustín Valverde	Universidad de Málaga, Spain

Organizing Commitee

Jesús Medina	Universidad de Cádiz, Spain
Valentín Liñeiro-Barea	Universidad de Cádiz, Spain
María Eugenia Cornejo-Piñero	Universidad de Cádiz, Spain
Juan Carlos Díaz-Moreno	Universidad de Cádiz, Spain
Ignacio García-García	Universidad de Cádiz, Spain
Eloísa Ramírez-Poussa	Universidad de Cádiz, Spain
María José Benítez-Caballero	Universidad de Cádiz, Spain

Sponsoring Institutions

Universidad de Cádiz, Spain
Széchenyi István University (Gyor)
Hungarian Fuzzy Association

Table of Contents

Preface

Mathematics is an indispensable field for a lot of areas such as Engineering, Computer Science, Physics, Chemistry and Business, in which improves the current methodologies and solves new challenges.

An important branch in Computer Science is Computational Intelligence, whose aim is to provide methods to deal with complex real-world problems for which traditional approaches are not feasible. Some of the methods that Computational Intelligence encompasses are, among others, fuzzy logic, evolutionary computation, neural networks, as well as probabilistic and statistical approaches, such as Bayesian networks or kernel methods.

In the past few years Computational Intelligence has become one of the main research topics at the Széchenyi István University. The first six Györ Symposia on Computational Intelligence have been successfully organized from 2008 to 2014. The seventh Györ Symposium on Computational Intelligence is jointly held with the fourth International Workshop on Mathematics and Soft Computing and it is called the 7th European Symposium on Computational Intelligence and Mathematics (ESCIM 2015). The location has been changed but preserves the philosophy of the past Györ Symposia enriching from a more mathematical perspective. That is, bringing together scientists and engineers working in the field of computational intelligence and mathematics to solve current challenges in these fundamental areas.

ESCIM 2015 will be held in Cádiz from October 7th to 10th, 2015, and it is organized by members of the University of Cádiz, Spain.

This symposium proceedings volume contains the contributions presented during ESCIM 2015, which have been included in different sections:

- Decision-Making under Uncertainty and Data Mining
- Evolutionary Computation, Metaheuristics and Machine Learning
- Software Verification and Validation
- Computational Optimization
- Computational Intelligence
- Mathematics and Soft Computing
- Formal Concept Analysis
- Graded Algebras and Algebras Admitting Multiplicative Bases
- Generalized Convexity and Fuzzy or Interval Valued Applications

We would like to thank the plenary speakers for their outstanding contributions to research and leadership in their respective fields. There were six plenary lectures covering the different areas of the symposium in charge of prestigious researches such as László Kóczy, Manuel Ojeda-Aciego, David Pearce, Jozef Pocs, Sandra Sandri and Nagy Szilvia.

We would also like to thank all the participants for their contributions to the symposium program and all the authors for their submitted papers. We are also indebted to the special session organizers and our colleagues members of the Program Committee, since the successful organization of this symposium would not have been possible without their work. Finally, we acknowledge the support received from the Department of Mathematics of the University of Cádiz, the Széchenyi István University (Györ) and the Hungarian Fuzzy Association.

October 2015

Jesús Medina
Conference Chair
ESCIM 2015

Convex Optimisation Problems in Bioengineering

Clemente Cobos Sanchez, Francisco Garcia-Pacheco, Jose Maria Guerrero Rodriguez, Angel Quiros Olozabal, and German Alvarez Tey

Depto. Ingeniería de Sistemas y Electrónica,
Depto. de Matemáticas Universidad de Cádiz, Cádiz, Spain
{clemente.cobos,garcia.pacheco,josem.guerrero,angel.quiros,german.alvarez}@uca.es

Abstract. *Many problems in engineering require to determine the spatial distribution of electric currents flowing on a conductive surface, which must satisfy some given requirements for the produced fields, electromagnetic energy, etc. The reconstruction of current distribution on the conducting surface subjected to these constraints is an inverse problem, which when formulated using boundary element methods can be posed as a convex optimisation. Here we present a convex optimisation framework to tackle problems in Bioengineering, that permits the prototyping of many different cost functions and constraints. Several examples of MRI gradients and TMS coils were designed and simulated to demonstrate the validity of the proposed approach.*

Keywords: Convex optimisation, Boundary element method, Bioengineering, Field synthesis

1 Introduction

Magnetic Resonance Imaging (MRI) has become an invaluable tool for diagnostic medicine. It is based on the use of well defined and controlled magnetic fields, as the magnetic field gradients, used to encode spatially the signals from the sample. These field gradients are generated by coils of wire, usually placed on cylindrical surfaces, although other geometries can be employed [1].

Transcranial Magnetic Stimulation (TMS) is a non-invasive technique to stimulate the brain [2], which is applied to studies of cortical effective connectivity, presurgical mapping, psychiatric and medical conditions, such as major depressive disorder, schizophrenia, bipolar depression, post-traumatic, stress disorder and obsessive-compulsive disorder, amongst others. In TMS, a strong, brief current pulse driven through a coil is used to induce an electric field stimulating neurons in the cortex.

The problem in MRI gradient coil and TMS coil design is to find optimal positions for the multiple windings of coils so as to produce fields with the desired spatial dependence and properties (low inductance, high gradient to current ratio, minimal resistance, good field gradient uniformity, high focality and field penetration depth, etc.). We refer the reader to [1]-[3] for a wider perspective on these topics.

TMS and MRI coil design are then two examples of problems in bioengineering where is required to determine the spatial distribution of electric currents flowing on a conductive surface, which satisfies given requirements for the produced fields, electromagnetic energy, etc.

An appropriate and realistic formulation of this type of problems can be achieved by using a boundary element method (BEM), and incorporating the idea of stream function. The current density in the surface is then a vector field that is piecewise uniform (see [4, 6]). By using this current model, electromagnetic inverse problems, such as MRI and TMS coil design, can be formulated as a constrained optimization.

In this work, we present a convex optimisation framework for the solution of electromagnetic inverse problems in Bioengineering, such as MRI gradient and TMS coil design problem, allowing the prototyping of many different cost functions and constraints. Two examples of gradient and TMS coils were designed and simulated to demonstrate this method, and prototypes coils were built and tested to validate it.

2 Physical Model

A model of the current under search can be achieved by using a boundary element method (BEM), that allows the current distribution to be defined in terms of the nodal values of the stream function and elements of the local geometry (see [4]).

So let us assume that the surface, S, on which we want to find the optimal current, is divided into T triangular elements with N nodes, which are lying at each vertex of the element. We can then note the vector containing the nodal values of the stream function as $\psi \in \mathbb{R}^N$, which is going to be the optimization variable in this work.

The use of this current model allows the discrete formulation of all the magnitudes and physical properties of the coil involved in the design. All problems tackled here are convex and can be generalised as

$$\begin{cases} \text{minimise } f_0(\psi) \\ \text{subject to } f_i(\psi) \leq b_i, \;\; 1 \leq i \leq m \end{cases} \tag{1}$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$ are convex functions for $i = 1, ..., m$.

3 MRI

3.1 Minimum inductance (stored energy) coil

The quality of the MRI images strongly depend on how linear the variation of the field is with position. Analogously, in order to improve the image formation process ideal coils should have a minimum inductance, which can be related to the stored magnetic energy and dictates the speed at which current can be put into the coil.

The problem of designing a MRI gradient coil with good field gradient uniformity and minimum inductance can be posed as [4]

$$\begin{cases} \text{minimise } \psi^T L \psi \\ \\ \text{subject to } \dfrac{\|B_z\psi - b_t\|_\infty}{\|b_t\|_\infty} \leq D_{max} \end{cases}$$

where

- $H \in \mathbb{N}$, with $N > H$, is the number of points where the target field is defined.
- $B_z \in \mathbb{R}^{H\times N}$ is a known matrix, where the coefficient $B_z(i,j)$ is the z-component of the magnetic induction produced by the current element associated to the j^{th}-node in the prescribed i^{th}-point.
- $b_t \in \mathbb{R}^H$ is the target field, prescribed in the H points.
- $L \in \mathbb{R}^{N\times N}$ is the inductance matrix, which is symmetric and positive-definite.
- The magnetic field is required to deviate by less than a given value from linearity, usually $D_{max} \sim 5\%$.

4 TMS

4.1 Minimum stored energy coil

Power requirements often limit the duration and frequency of repetitive TMS, for example, via coil heating. Thus, an ideal TMS coil should produce a strong stimulation of a prescribed region, and a minimum electric field in the rest of non target regions; and have a minimum stored magnetic energy.

The problem of designing a TMS coil that produces a maximum field in a target region with minimum stored energy can be posed as

$$\begin{cases} \text{maximise } \|B\psi\|_2 \\ \text{minimise } \psi^T L \psi \end{cases} \tag{2}$$

where

- $H \in \mathbb{N}$, with $N > H$, is the number of points where the target field is defined.
- $B \in \mathbb{R}^{H\times N}$ is a known matrix, where the coefficient $B(i,j)$ is the modulus of the magnetic induction produced by the current element associated to the j^{th}-node in the prescribed i^{th}-point.
- $b_t \in \mathbb{R}^H$ is the target field prescribed in the H points, which in TMS can be considered constant, that is, $b_t(i) = b_t(j)$ for all $i, j = 1, ..., H$.
- $L \in \mathbb{R}^{N\times N}$ is the inductance matrix, which is symmetric and positive-definite.

This type of problem is also known as Tikhonov regularised minimisation. According to Subsection 5.6 and *1* of Corollary 1, this problem can be equivalently written as

$$\begin{cases} \text{maximise} \left\| \left(BC^{-1}\right) \psi \right\|_2 \\ \text{subject to} \left\| \psi \right\|_2 = 1 \end{cases}$$

where $L = C^T C$ is the Cholesky decomposition of the inductance matrix L (which is symmetric and positive-definite).

Other admissible reformulations (see *2* of Corollary 1) are

$$\begin{cases} \text{minimise} \left\| \psi \right\|_2 \\ \text{subject to} \left\| \left(BC^{-1}\right) \psi \right\|_2 = \left\| \left(BC^{-1}\right) \right\|_2 \end{cases}$$

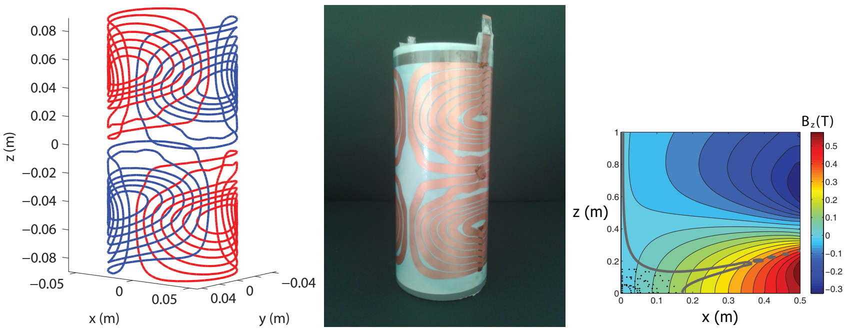
or more generally

$$\begin{cases} \text{minimise} \left\| \psi \right\|_2 \\ \text{subject to} \left\| \left(BC^{-1}\right) \psi - b_t \right\|_2 = \left\| \left(BC^{-1}\right) \right\|_2 \end{cases}$$

for b_t constant.

5 Results

5.1 MRI



(a) (b) (c)

Fig. 1. a) equivalent numerical model of coil in Fig. 5.1 (red and blue colors are used to indicate wires in which there is a different sense of current flow). b) Photograph of the constructed prototype coil. c) Contours of the B_z field produced by the wire arrangement Fig 5.1. The grey line delineates the region where the field deviates by less than 5% from linearity.

A prototype cylindrical transverse MRI gradient coil has been designed using the proposed convex framework (Fig. 5.1), it corresponds to a minimum inductance transverse cylindrical coil of radius 4.5 cm and height 18 cm, designed to produce a field gradient which deviates from linearity by less than 5% within a central, uniform distribution of 400 points spread over a sphere of radius of 3.5 cm. This prototype coil has been constructed using a variable track width produced in a flexible printed circuit board (PCB), where the copper thickness was 35 μm. The former on which the prototype tracks were laid was a cylinder of polyvinylchloride (PVC) with 3.4 mm thickness.

Figure 5.1 shows the gradient coil prototype, that when was connected to 6.0 A DC current supply (Agilent U8031A, USA), produces the magnetic field in Fig. 5.1. This coil when energized produce the B_z field displayed in Fig. 5.1, which satisfies the initial requirements, as we can see the target region is within the grey line that delineates the volume where the field deviates by less than 5% from linearity.

5.2 TMS

The proposed approach has also been used to produce a TMS stimulator on a rectangular former of dimensions 20 cm × 10 cm, designed to have minimum stored energy and to maximize the magnetic field in a prescribed spherical volume of interest of radius 2.0 cm that is centred 4.0 cm below the center of the coil plane, as shown in Fig. 5.2 where the coil solution is also depicted.

In order to validate the proposed solution, we manufactured the corresponding prototype coil, which was wound with 1.5 mm thick continuous copper wire, Fig. 5.2. This TMS prototype when connected to 6.0 A DC current supply (Agilent U8031A, USA), produces the magnetic field in Fig. 5.2, which was measured using a magnetic flow sensor MAG 3100. The results obtained here indicates that the prototype TMS coil produces a remarkably high magnetic field in the target volume that decreases rapidly out of the volume.

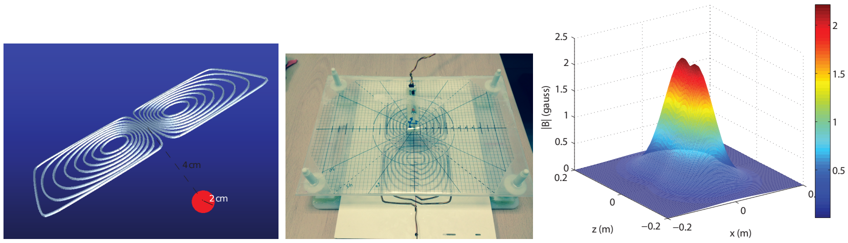
5.3 Numerical Implementation

Software was written in Fortran 90 to tackle the problems presented here, and the produced optimal solutions were found in good agreement with those generated using software for convex programming such us CVX [7] The Fortran 90 software also includes subroutine that allows the testing of the coil designs, as it calculates the field produced by the wire pattern via Biot-Savart integration.

5.4 Conclusion

Here we present a convex optimisation framework to tackle problems in Bioengineering. This is a powerful approach for designing of MRI gradient, and novel method for the generation of TMS coils wounds on arbitrarily shaped surface.

The method has been experimentally validated by constructing and testing prototype coils, where the magnetic fields produced show the accuracy of the proposed technique.



(a) (b) (c)

Fig. 2. a) Schematic diagram showing the TMS coil solution and along with the region of interest in which the desired magnetic field must be maximized. b) Photograph of the constructed TMS prototype coil and the experimental set up to measure the magnetic field with the flow sensor MAG 3100. c) Magnetic field modulus in a 20 cm × 20 cm plane which is centred 4.0 cm below the center of the coil plane.)

References

1. Turner, R.: Gradient coil design: A review of methods, Magn. Reson. Imaging, 11, 903 – 920(1993).
2. Wassermann E. M., Epstein C. M., Ziemann U., Walsh V., Paus T., Lisanby S. H., editors. The Oxford handbook of transcranial magnetic stimulation. New York: Oxford University Press; 2008
3. Koponen L. M., Nieminen J. O., Ilmoniemi R.J. Minimum-energy coils for transcranial magnetic stimulation: application to focal stimulation. Brain Stimul 2014.
4. Cobos Sanchez, C., S. G. Garcia, L. D. Angulo, C. M. De Jong Van Coevorden, and A. Rubio Bretones. A divergence-free BEM method to model quasi-static currents: Application to MRI coil design. Progress In Electromagnetics Research B, Vol. 20, 187–203, 2010.
5. Poole M., Weiss P., Lopez H.S., Ng M., Crozier S., Minimax current density coil design, J. Phys. D: Appl. Phys. 43 (9) (2010) 095001. 13pp.
6. Poole M., Jon Shah N. Convex optimisation of gradient and shim coil winding patterns. Journal of Magnetic Resonance, Volume 244, 2014, Pages 36-45.
7. CVX Research, Inc., CVX: Matlab Software for Disciplined Convex Programming, version 2.0, (August 2012).

Appendix: Mathematical foundations of the previous models

This appendix is devoted to provide the mathematical foundations needed to express the previous problems in the form given in Equation (1).

5.5 Brief introduction and background

Let A be an $H \times N$ matrix and consider a norm $\|\cdot\|$ in $\mathbb{R}^N$. For every $\varepsilon \geq 0$ we can consider the closed $\|\cdot\|$-ball of center 0 and radius ε, which is a compact subset of $\mathbb{R}^N$:

$$\mathsf{B}_{\|\cdot\|}(0,\varepsilon) = \{\psi \in \mathbb{R}^N : \|\psi\| \leq \varepsilon\}.$$

When $\varepsilon = 1$ then we will simply write $\mathsf{B}_{\|\cdot\|}$ in lieu of $\mathsf{B}_{\|\cdot\|}(0,1)$. It is obvious that $\mathsf{B}_{\|\cdot\|}(0,\varepsilon) = \varepsilon \mathsf{B}_{\|\cdot\|}$ for all $\epsilon \geq 0$.

It can be proved that the sup of A on the previous ball is attained at an element of its sphere. In other words,

$$\max\{\|A\psi\| : \|\psi\| \leq \varepsilon\} = \max\{\|A\psi\| : \|\psi\| = \varepsilon\} = \sup\{\|A\psi\| : \|\psi\| < \varepsilon\}.$$

We will denote by $\exp(A,\varepsilon)$ to the set of all those elements of the sphere of the ball above at which the previous max is attained. In other words,

$$\exp_{\|\cdot\|}(A,\varepsilon) := \left\{\varphi \in \mathsf{B}_{\|\cdot\|}(0,\varepsilon) : \|A\varphi\| = \max\{\|A\psi\| : \|\psi\| \leq \varepsilon\}\right\}.$$

Again, when $\varepsilon = 1$ we will write $\exp_{\|\cdot\|}(A)$ instead of $\exp_{\|\cdot\|}(A,1)$. On the other hand, note that $\exp_{\|\cdot\|}(A,\varepsilon) = \varepsilon \exp_{\|\cdot\|}(A)$ for all $\varepsilon \geq 0$.

The norm of the matrix A is by definition $\|A\| := \max\{\|A\psi\| : \|\psi\| \leq 1\} = \max\{\|A\psi\| : \|\psi\| = 1\}$ and thus the elements of $\exp_{\|\cdot\|}(A)$ are precisely the vectors of $\mathbb{R}^N$ at which A attains its norm (these vectors will be called the supporting vectors of A). It can be proved that $\|A\chi\| \leq \|A\|\|\chi\|$ for all $\chi \in \mathbb{R}^N$.

Recall that a (real) scalar product on $\mathbb{R}^N$ is defined by a positive-definite symmetric matrix P as $(\varphi,\psi) := \varphi P \psi$. This scalar product makes $\mathbb{R}^N$ a Hilbert space whose norm is $\|\varphi\|_P := (\varphi,\varphi)^{\frac{1}{2}}$. It is well known among the functional analysts that all the Hilbert spaces of the same dimension are linearly isometric, which means that there exists a surjective linear isometry $T_P : \left(\mathbb{R}^N, \|\cdot\|_P\right) \to \left(\mathbb{R}^N, \|\cdot\|_2\right)$. If we keep denoting by T_P to the matrix associated to the isometry T_P, then $\|T_P\psi\|_2 = \|\psi\|_P$ for all $\psi \in \mathbb{R}^N$. In matrix theory, the expression of T_P is given by the Cholesky decomposition of the symmetric positive-definite matrix P, that is, $P = L^T L$. Indeed, notice that

$$\|T_P\psi\|_2^2 = \|\psi\|_P^2 = \psi^T P \psi = \psi^T (L^T L)\psi = (L\psi)^T (L\psi) = \|L\psi\|_2^2$$

which means that we can take $T_P := L$.

5.6 Uniformizing norms

Consider the optimization problem given in Equation 2

$$\begin{cases} \max \|B\psi\|_2 \\ \min \psi^T L \psi \end{cases} \quad \psi \in \mathbb{R}^N$$

If we assume that L is symmetric and positive-definite, then there exists a matrix T_L such that $\psi^T L \psi = \|T_L\psi\|_2$. Therefore, the above problem becomes

$$\begin{cases} \max \|B\psi\|_2 \\ \min \|T_L\psi\|_2 \end{cases} \quad \psi \in \mathbb{R}^N$$

Taking into consideration that T_L is invertible, we can rewrite it as

$$\begin{cases} \max \|A\varphi\|_2 \\ \min \|\varphi\|_2 \end{cases} \varphi \in \mathbb{R}^N$$

where $A := BT_L^{-1}$.

5.7 Turning maximization problems into minimization problems (without losing convexity)

The following results throws some light on the problem described in Equation (2) in the sense that conjugation of the two conditions of maximization and minimization makes it an unsolvable problem.

Theorem 1. *Let A be an $H \times N$ matrix and consider a norm $\|\cdot\|$ in $\mathbb{R}^N$. Consider the optimization problem*

$$\begin{cases} \max \|A\varphi\| \\ \min \|\varphi\| \end{cases} \varphi \in \mathbb{R}^N$$

1. *There does not exist $\varphi \in \mathbb{R}^N$ such that, for all $\psi \in \mathbb{R}^N$, $\|A\varphi\| \geq \|A\psi\|$ and $\|\varphi\| \leq \|\psi\|$.*
2. *There are infinitely many $\varphi \in \mathbb{R}^N$ such that, for all $\psi \in \mathbb{R}^N$, either $\|A\varphi\| \geq \|A\psi\|$ or $\|\varphi\| \leq \|\psi\|$. These solutions are the elements of the set*

$$\bigcup_{\varepsilon \geq 0} \exp_{\|\cdot\|}(A, \varepsilon).$$

Proof.

1. Suppose to the contrary that there is $\varphi \in \mathbb{R}^N$ such that $\|A\varphi\| \geq \|A\psi\|$ and $\|\varphi\| \leq \|\psi\|$ for all $\psi \in \mathbb{R}^N$. Since $\|\varphi\| \leq \|\psi\|$ for all $\psi \in \mathbb{R}^N$ we must have that $\varphi = 0$ which then contradicts that $\|A\varphi\| \geq \|A\psi\|$ for all $\psi \in \mathbb{R}^N$ since $A\varphi = 0$.
2. In the first place, assume that there exists $\varphi \in \mathbb{R}^N$ such that $\|A\varphi\| \geq \|A\psi\|$ or $\|\varphi\| \leq \|\psi\|$ for all $\psi \in \mathbb{R}^N$. We will show that $\varphi \in \exp_{\|\cdot\|}(A, \varepsilon)$ for $\varepsilon := \|\varphi\|$. Indeed, let $\psi \in \mathbb{R}^N$ such that $\|\psi\| < \varepsilon \, (= \|\varphi\|)$. Then by assumption $\|A\varphi\| \geq \|A\psi\|$, which means that $\|A\varphi\| \geq \sup\{\|A\psi\| : \|\psi\| < \varepsilon\} = \max\{\|A\psi\| : \|\psi\| \leq \varepsilon\} \geq \|A\varphi\|$ since the open ball $\{\psi \in \mathbb{R}^N : \|\psi\| < \varepsilon\}$ is dense in the closed ball $\{\psi \in \mathbb{R}^N : \|\psi\| \leq \varepsilon\}$. As a consequence, $\varphi \in \exp_{\|\cdot\|}(A, \varepsilon)$.
   Conversely, we will show that every $\varphi \in \exp_{\|\cdot\|}(A, \varepsilon)$ verifies that $\|A\varphi\| \geq \|A\psi\|$ or $\|\varphi\| \leq \|\psi\|$ for all $\psi \in \mathbb{R}^N$. Indeed, fix any $\psi \in \mathbb{R}^N$. If $\varepsilon \leq \|\psi\|$, then we are done because $\|\varphi\| = \varepsilon$. If $\varepsilon > \|\psi\|$, then we conclude that

$$\|A\varphi\| = \max\{\|A\chi\| : \|\chi\| \leq \varepsilon\} \geq \|A\psi\|.$$

Corollary 1. *Let A be an $H \times N$ matrix and consider a norm $\|\cdot\|$ in $\mathbb{R}^N$. The optimization problem*

$$\begin{cases} \max \|A\varphi\| \\ \min \|\varphi\| \end{cases} \varphi \in \mathbb{R}^N$$

is equivalent to any one of the following:

1. *The optimization problem*

$$\begin{cases} \max \|A\varphi\| \\ \|\varphi\| = 1 \end{cases} \varphi \in \mathbb{R}^N$$

which consists of finding the elements of $\exp_{\|\cdot\|}(A)$*, that is, the elements of* $\mathbb{R}^N$ *at which* A *attains its norm.*

2. *The convex optimization problem*

$$\begin{cases} \min \|\varphi\| \\ \|A\varphi\| = \|A\| \end{cases} \varphi \in \mathbb{R}^N$$

which again consists of finding the supporting vectors of A*.*

Proof.

1. In accordance to Theorem 1, the solutions to the optimization problem

$$\begin{cases} \max \|A\varphi\| \\ \min \|\varphi\| \end{cases} \varphi \in \mathbb{R}^N$$

are the elements of the sets

$$\bigcup_{\varepsilon \geq 0} \exp_{\|\cdot\|}(A, \varepsilon).$$

Since $\exp_{\|\cdot\|}(A, \varepsilon) = \varepsilon \exp_{\|\cdot\|}(A)$ for all $\varepsilon > 0$, it suffices to find the supporting vectors of A, that is, the elements of $\exp_{\|\cdot\|}(A)$, which are precisely the solutions to the

$$\begin{cases} \max \|A\varphi\| \\ \|\varphi\| = 1 \end{cases} \varphi \in \mathbb{R}^N$$

2. All we need to show is that the solutions to the convex optimization problem

$$\begin{cases} \min \|\varphi\| \\ \|A\varphi\| = \|A\| \end{cases} \varphi \in \mathbb{R}^N$$

are the supporting vectors of A. Indeed, let $\varphi \in \exp_{\|\cdot\|}(A)$. If $\psi \in \mathbb{R}^N$ and $\|A\psi\| = \|A\|$, then we have that

$$\|\varphi\| = 1 = \frac{\|A\psi\|}{\|A\|} \leq \frac{\|A\|\|\psi\|}{\|A\|} = \|\psi\|$$

which implies that φ is a solution of the convex minimization problem. Conversely, let φ a solution of the convex minimization problem. We will prove that $\varphi \in \exp_{\|\cdot\|}(A)$. By assumption, $\|A\varphi\| = \|A\|$ so all we need to show is that $\|\varphi\| = 1$. Indeed, it suffices to consider any $\psi \in \exp_{\|\cdot\|}(A)$. Since $\|A\psi\| = \|A\|$ we have that $\|\varphi\| \leq \|\psi\| = 1$. The other inequality follows from the fact that $\|A\| = \|A\varphi\| \leq \|A\|\|\varphi\|$.

Formalization in ACL2 of Matrix Algebra Basic Concepts*

J.L. Pro–Martín , J.L. Ruiz–Reina and F.J. Martín–Mateos
jlpro@modinem.com, fjesus@us.es, jruiz@us.es

Computational Logic Group
Dept. of Computer Science and Artificial Intelligence, University of Seville
E.T.S.I. Informática, Avda. Reina Mercedes, s/n. 41012 Sevilla, Spain

Abstract. In this paper we present a formalization of basic operations on matrices in the ACL2 theorem prover, including addition, product, transpose, and inverse of matrices. We define these operations and give proofs of their main properties. The main result is an ACL2 implementation and formal verification of the Gauss-Jordan algorithm for computing the inverse of a matrix. Our formalization is based on *abstract stobjs*, an ACL2 feature allowing both convenient logical reasoning and execution efficiency. In fact, we get quite good execution time response for big matrices of several hundred thousands elements. The complete formalization can be found in [1].

1 Formalizing matrices in ACL2

The ACL2 system [4] is both a programming language, a logic for reasoning about programs defined in the language, and a mechanical theorem prover to assist in the proof process. The programming language is an extension of an applicative subset of Common Lisp and the logic is quantifier-free, first-order with equality, including a rule of inference of proof by structural induction. See [2] for a detailed description of ACL2.

ACL2 has been widely used in formal validation of hardware algorithms and in many math fields. As for matrix algebra, a previous work is [3], where a formalization of the Strassen algorithm for matrix multiplication is presented, using applicative *record structures*. Also, in [5] a number of operations (including inverse) are defined and some of their properties proved, using ACL2 bidimensional arrays. Since one of our main concerns is to improve execution efficiency, it is tempting to apply some interesting ACL2 features designed for that purpose, in a formalization of matrix algebra. Among these features are *stobjs* (for *Single Threaded OBJects*) and *abstract stobjs*.

In principle, if we use a list-based representation of matrices, accessing and updating cannot be done in constant time, because we need copying, as usual in an applicative setting. This feature is critical if we want to define algorithms

* This work was partially supported TIN2013-41086-P project (Spanish Ministry of Economy and Competitiveness), cofinanced with FEDER funds

dealing with big matrices. Fortunately, ACL2 stobjs allow lookups and (destructive) updates in constant time. When an object is declared to be single-threaded, ACL2 enforces certain syntactic restrictions on its use, ensuring that only one copy of the object is ever needed. In this way, this efficient data structure is consistent with the applicative semantics of ACL2.

Stobjs are structures composed by fields. These fields can store a single data or *unidimensional* arrays. But a matrix is intuitively represented by a two-dimensional array. For example, element a_{ij} of matrix A maps to element of the ith-row and jth-column. This use of two indices to access matrix elements is so extended in mathematics that we should preserve it in our formalization.

But here arises the very first problem: currently *stobj's* only allow one-dimensional arrays, so we define this stobj in ACL2 (the suffix "`...$c`" stands for *concrete* and will be explained later):

```
(defstobj matrix$c
 (m$c        :type (array rational (1)) :initially 0 :resizable t)
 (nrows$c    :type (integer 1 *) :initially 1)
 (ncolumns$c :type (integer 1 *) :initially 1))
```

Where `m$c` is the one-dimensional array that supports matrix elements, `nrows$c` and `ncolumns$c` gives us the number of rows and columns of the matrix defined above.

From the `:LOGIC` point of view a *stobj* will be represented by means of a list (made with *conses*) of elements. Also an array field of an stobj, from the logic point of view, is represented as a list of elements. For example, if we have A the following 2×3 matrix, in the logic we would have the representation:

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \longrightarrow \texttt{((1 2 3 4 5 6) 2 3)}$$

If we denote c the number of columns of A, r the number of rows of A and B the contents of the array field, it's easy to see the following mappings between elements in both representations:

1. $a_{ij} \mapsto b_{i \cdot c+j}$
2. $b_k \mapsto a_{\lfloor k/r \rfloor, k \mod r}$

So in the `:LOGIC` we would have to reason with this additional disadvantage of dealing with these mappings. This would increase the difficulty of our proofs and decrease their clarity. Fortunately, abstract stobjs allows us to use *stobj* in the `:EXEC` world but two-dimensional access in `:LOGIC` world.

1.1 *Abstract stobj's* in ACL2

Abstract stobj's (as opposed to *concrete stobj's*) is a relatively new ACL2 feature introduced in [6]. An abstract stobj provides an alternative logical representation of a concrete stobj. That is, we can define a *simpler* logical representation of the concrete *stobj* in order to abstract its complexity in terms of reasoning. In our case, we will be able to give a natural and two-dimensional access matrix representation in the `:LOGIC` and a more complex but efficient one-dimensional array representation in the `:EXEC` for our matrices.

We introduce an *abstract stobj* using the `defabsstobj` event. Two functions are needed in this event: a *recognizer* and a *correspondence* function. In order to be finally admitted, some proof-obligations are generated:

1. `:correspondence`. Theorems establishing that abstract and concrete stobj's represent the same inner values in the terms of the correspondence function.
2. `:preserved`. Theorems that prove that updaters maintains the recognizer function.
3. `:guard-thm`. Theorem proving that guards are verified in every call in the `:EXEC` world.

These proof-obligations have to be proved once and for all, ensuring that the logical correspondence between the abstract stobj and its associated concrete stobj is preserved. So we can define the following abstract stobj:

```
(defabsstobj matrix
    :concrete matrix$c
    :recognizer (matrixp    :logic matrix$ap :exec matrix$cp)
    :creator (create-matrix :logic create-matrix$a
                            :exec create-matrix$c)
    :corr-fn matrix$corr
    :exports ((nrows    :logic nrows$a    :exec nrows$c)
              (ncolumns :logic ncolumns$a :exec ncolumns$c)
              (lookup   :logic lookup$a   :exec lookup$c)
              (update   :logic update$a   :exec update$c)
              (redim    :logic redim$a    :exec redim$c)))
```

Where the suffix "`...$a`" stands now for *abstract*. We can see here the primitives defined for our matrix object where their names (`nrows`, `ncolumns`, and so on) are some kind of self-explanatory about its behaviour. Let's see with some detail the function `(lookup A i j)`, that returns the a_{ij} element. We must define the `:EXEC` and `:LOGIC` version of `lookup`:

```
(defun lookup$c (matrix$c i j)
    (m$ci (+ (* i (ncolumns$c matrix$c)) j) matrix$c))

(defun lookup$a (matrix$a i j)
    (nth j (nth i matrix$a)))
```

So to establish that `lookp$c` and `lookup$a` return the same value we must prove the following theorem in ACL2 (we have removed some not relevant conditions in the clause for the sake of clarity):

```
(defthm lookup{correspondence}
    (implies (matrix$corr matrix$c matrix)
             (equal (lookup$c matrix$c i j)
                    (lookup$a matrix i j))))
```

This theorem can be read as follows: “if matrix$c and matrix correspond to each other, lookup will return the same value”. Once this is proved, we can use the primitives with the desired logic interface. The rest of the functions that use matrices must do it by means of those primitives.

2 Defining operations over matrices

We have defined some operations over matrices using, as explained, the given primitives in the defabsstobj event described above. Experience in the use of ACL2 prover tells us that the way one defines functions can dramatically change proofs complexity, so we tried to define functions in such a simple way that principle of induction could perform quite well.

For example, we can define the matrix addition like this:

```
(defun add-matrix-row (A B m n)
    (if (zp n)
        (update A m 0 (+ ( lookup A m 0) ( lookup B m 0)))
        (seq A
            (update A m n (+ (lookup A m n)
                             (lookup B m n)))
            (add-matrix-row A B m (1- n)))))

(defun add-matrix-rows (A B m n)
    (if (zp m)
        (add-matrix-row A B 0 n)
        (seq A
            (add-matrix-row A B m n)
            (add-matrix-rows A B (1- m) n))))

(defun add-matrix (A B)
    (add-matrix-rows A B (1- (nrows A))
                         (1- (ncolumns A))))
```

So, add-matrix-row adds the elements of m row, from 0 to n column. The next function, add-matrix-rows, note the plural, adds rows from 0 to m, and each of this, from column 0 to n. Finally, we have add-matrix function, that only makes first calling to last function with proper arguments to start the computation. This way of defining functions can be considered as a pattern that can be applied to other functions such as matrix equality, transposed matrix, matrix product and so on.

It turns out that this recursive pattern is specially well-suited when we prove properties of these functions by induction. That is, first we prove the property on only a given row of the matrix, by induction in the number of columns. Afterwards we prove it in a subset of rows of the matrix, to finally be able to prove it for the whole matrix. It is worth mentioning that almost every property (with some hard exceptions) can be proved in this way.

2.1 List of some proved properties

We now list some of the main properties proved in our formalization. For details, see the whole formalization in [1].

Transposition

$(A^T)^T = A$

$I_n^T = I_n$

Opposite

$-(-A) = A$

$(-A)^T = -(A^T)$

Scalar product

$\alpha \cdot (\beta \cdot A) = (\alpha \cdot \beta) \cdot A$

$0 \cdot A = \emptyset$

$\alpha \cdot \emptyset = \emptyset$

$1 \cdot A = A$

$-1 \cdot A = -A$

$(\alpha \cdot A)^T = \alpha \cdot A^T$

$(\alpha \cdot A)^T = \alpha \cdot A^T$

$\alpha \cdot (-A) = -(\alpha \cdot A)$

Addition

$A + B = B + A$

$(A+B)+C = A+(B+C)$

$A + \emptyset = A$

$\emptyset + A = A$

$A + (-A) = \emptyset$

$-(A + B) = -A + (-B)$

$(\alpha + \beta) \cdot A = \alpha A + \beta A$

$\alpha(A + B) = \alpha A + \alpha B$

$A + A = 2A$

$(A + B)^T = A^T + B^T$

$A + B = \emptyset \rightarrow A = -B$

$-A + A = \emptyset$

Product

$A \cdot \emptyset = \emptyset$

$\emptyset \cdot A = \emptyset$

$I_n \cdot A = A$

$A \cdot I_n = A$

$(A \cdot B) \cdot C = A \cdot (B \cdot C)$

$A \cdot (B + C) = A \cdot B + A \cdot C$

$(A + B) \cdot C = A \cdot C + B \cdot C$

$-A \cdot B = -(A \cdot B)$

$A \cdot (-B) = -(A \cdot B)$

$A \cdot (\alpha \cdot B) = \alpha \cdot (A \cdot B)$

$(\alpha \cdot A) \cdot B = \alpha \cdot (A \cdot B)$

$(A \cdot B)^T = B^T \cdot A^T$

Row operations

$F_{ii}A = A$

$F_{ij}I_n \cdot A = F_{ij}A$

$F_{ij}B \cdot A = F_{ij}(B \cdot A)$

$F_i(\alpha)I_n \cdot A = F_i(\alpha)A$

$F_i(\alpha)B \cdot A = F_i(\alpha)(B \cdot A)$

$F_{ij}(\alpha)I_n \cdot A = F_{ij}(\alpha)A$

$F_{ij}(\alpha)B \cdot A = F_{ij}(\alpha)(B \cdot A)$

Where:

- A, B, C: Matrices of arbitrary dimensions.
- α, β: Arbitrary scalars.
- $\emptyset$: Matrix of arbitrary dimensions where all the elements are 0.
- I_n: The nth-order identity matrix.
- $F_{ij}A$: Swaps rows i and j of matrix A.
- $F_i(\alpha)A$: Multiply i row of matrix A by α.
- $F_{ij}(\alpha)A$: Adds, to i row of matrix A, the j row multiplied by α.

2.2 Gauss-Jordan algorithm implementation

Using the above row transformations F_{ij}, $F_i(\alpha)$ and $F_{ij}(\alpha)$, we can implement the Gauss-Jordan algorithm to get the inverse of a matrix. The main idea is that we begin with the (composed) matrix $(I_n|A)$ and we apply a sequence of row transformations, trying to transform the matrix A to I_n. We have proved that if the final result of these operations is the composed matrix $(B|C)$ and $C = I_n$, then $B \cdot A = I_n$, and, therefore, $B = A^{-1}$. That is we have defined and formally verified a Gauss-Jordan based algorithm for computing the inverse of a matrix.

Abstract stobjs turns out to be crucial for the whole formalization. They allow us to reason about the algorithm as if it were executed on a bidimensional and intuitive representation of matrices, although actually is executed using the more efficient (but more complex) unidimensional, stobj based, representation.

2.3 Execution efficiency

We have tested our implementation computing the main operations (addition, multiplication and inverse) on randomly generated matrices. For example, we compute the sum of matrices of dimensions up to $n = 1000$, product for dimensions up to $n = 300$ and inverse of matrices of dimensions up to $n = 100$. We compared our execution time with the execution time obtained using the implementation in [5]. We also compared memory allocation in both implementations.

Our implementation outperforms the implementation in [5] in all cases. In the case of addition and product, our implementation is more than three times faster. In the case of the inverse, although the execution times are more similar, our implementation is still about 20% faster. In the next tables we can see measured execution times for sum (left) and product (right) of matrices (all times in seconds):

Dimension	**Stobj's**	**Gamboa**
100	0,00	0,01
200	0,00	0,04
300	0,01	0,20
400	0,01	0,26
500	0,01	0,35
600	0,02	0,66
700	0,02	0,83
800	0,03	1,34
900	0,03	1,93
1000	0,04	2,67

Dimension	**Stobj's**	**Gamboa**
30	0,00	0,01
60	0,01	0,03
90	0,03	0,06
120	0,07	0,16
150	0,10	0,25
180	0,18	0,43
210	0,28	0,69
240	0,39	1,05
270	0,56	1,60
300	0,78	2,01

References

1. `http://www.glc.us.es/acl2/matrix/acl2-matrix.tgz`.
2. M. Kaufmann, P. Manolios, and J S. Moore. *Computer-Aided Reasoning: An Approach.* Kluwer Academic Publishers, 2000.
3. F. Palomo, I. Medina and J.A. Alonso. Certification of Matrix Multiplication Algorithms (Strassen's Algorithm in ACL2) In *TPHOLs 2001.*
4. M. Kaufmann and J S. Moore. *ACL2 Version 7.1*, 2015.
   Homepage: `http://www.cs.utexas.edu/users/moore/acl2/`
5. J. Cowles, R. Gamboa and J.V. Baalen. Using ACL2 arrays to formalize matrix algebra. In *ACL2Workshop 2003.*
6. S. Goel, W. Hunt, and M. Kaufmann. Abstract Stobjs and Their Application to ISA Modeling. In *R. Gamboa and J. Davis (Eds.): ACL2 Workshop 2013*, pp 54-69.

Preference-Based Genetic Algorithm for Solving the Bio-Inspired NK Landscape Benchmark

Christof Ferreira Torres, Sune S. Nielsen, Grégoire Danoy, and Pascal Bouvry

FSTC, University of Luxembourg
6, rue Richard Coudenhove-Kalergi,
Kirchberg, Luxembourg
christof.ferreira.001@student.uni.lu
{sune.nielsen,gregoire.danoy,pascal.bouvry}@uni.lu

Abstract. In molecular biology, the subject of protein structure prediction is of continued interest, not only to chart the molecular map of living cells, but also to design proteins with new functions. In this work a Preference-Based Genetic Algorithm (PBGA) is proposed aiming to optimise NK Landscape based benchmarks designed and shown to mimic the properties of the protein Inverse Folding Problem (IFP). The PBGA algorithm incorporates a weighted sum model in order to combine fitness and diversity into a single objective function scoring a set of individuals as a whole. By adjusting the sum weights, a direct control of the fitness vs. diversity trade-off in the algorithm population is achieved by means of a selection scheme iteratively removing the least contributing individuals. Experimental results demonstrate the better performance of the PBGA algorithm compared to other state-of-the-art algorithms both in terms of fitness and diversity.

1 Introduction

Protein engineering in general aims at designing molecules with desired properties. A method allowing to successfully design such molecules would find applications in a number of areas such as designing improved enzymes for biotechnology applications or new antibodies towards already known targets. However evaluating and therefore optimising real biological instances is very computationally demanding. A novel approach recently proposed by Nielsen et al. consists in an NK Landscape benchmark suite that mimics the properties of the Inverse Folding Problem (IFP) [6]. The IFP aims, given a protein sequence of N amino acids, at finding other sequences that will result in the same 3D structure. The corresponding optimisation problem is highly *multi-modal* and the genetic algorithm proposed in this work addresses this aspect by adding a novel diversity controlling mechanism. The preference-based approach employs a Weighted Sum Model (WSM) in order to control the desired bias between fitness and diversity. The resulting WSM score allows to iteratively determine and remove the individual in the combined parent and offspring population, with the lowest overall fitness contribution with respect to the defined preferences. The remainder of this article is organised as follows. First the current state-of-the-art is situated in the related literature in Section 2, then a detailed description of the problem and of the biological background is introduced in Section 3. In Section 4 the contribution of this work in terms of achieving an adjustable level of fitness and diversity as a Preference-Based Genetic Algorithm (PBGA) is presented. Section 5 describes the experiments conducted and provides the analysis of the results obtained for the NK benchmark suite. Finally the conclusion and perspectives are summarised in Section 6.

2 State-of-the-art

In meta-heuristics, the subject of exploration vs. exploitation characteristics has been thoroughly studied. In this aim, a number of works have sought to maintain and control diversity in population-based meta-heuristics, e.g. crowding methods by DeJong [2], fitness sharing by Goldberg and Richardson [3], cellular algorithms by Alba and Dorronsoro [1], diversity preserving selection strategies based on hamming distance Shimodaira [7] and on altruism by Laredo et al. [4].

Preference-based algorithms have been discussed in the literature [5, 8] and refer to algorithms where the user preference is incorporated in the choice of regions in the solution or objective space. Preference can be incorporated in a number of ways, e.g., by modifying the fitness evaluation or selection schemes. The Indicator Based Evolutionary Algorithm (IBEA) [9] is an example where an indicator that characterises the population as a whole is used to guide the algorithm by eliminating the least desired individuals of the parent and offspring population union. The proposed PBGA in this paper uses the same principle of iterative elimination, determining the overall most preferable subset directly rather than achieving it as an indirect effect of designed mechanisms.

3 Bio-Inspired NK Landscape Benchmark Problem

In the NK benchmark problem as well as in the Inverse Folding Problem (IFP), a single solution is represented as a sequence $A = \{aa_i\}$ and consists of N residue positions, where $1 \leq i \leq N$ and $aa_i \in \{1, ..., 20\}$ corresponds to the set of 20 possible amino acids. The overall size and the number of local "hills and valleys" of the NK landscape model can be adjusted with two parameters, N and K. This paper focuses on optimising two novel NK benchmark model instances[1] proposed by Nielsen et al. [6]. These consist in the combination of two NK models, $F^A(x)$ and $F^B(x)$, by a simple multiplication with different K and different neighbourhood definitions as defined in the Table 1.

[1] The NK Landscape Protein IFP Benchmark Suite - `http://nk-ifp-bench.gforge.uni.lu/index.html`

Table 1: NK Landscape Protein IFP Benchmarks

Model	**Setting**
NK-IFP-1	$F^A(x)$: a $K = 4$ semi-adjacent circular neighbourhood is designed as follows: $\{x_{i2}, x_{i1}, x_{i+1}, x_{i+2}\}$, omitting the central position x_i. $F^B(x)$: a $K = 3$ neighbourhood of uniform random distribution.
NK-IFP-2	$F^A(x)$: a $K = 4$ semi-adjacent circular neighbourhood as *NK-IFP-1.* $F^B(x)$: a $K = 5$ neighbourhood of uniform random + 20 positions wide triangular distribution.

4 A Novel Preference-Based Approach

The main idea of the preference-based approach is to use a Weighted Sum Model (WSM) in order to constantly maintain a current population best fulfilling the defined preferences. In an iterative manner, the weakest individuals from the combination of parent and offspring populations are determined and removed until the desired population size is achieved.

Algorithm 1 Preference-Based Genetic Algorithm

```
Initialise($P_0$)
$t \leftarrow 0$
while $t < t_{max}$ do
    $Q_t \leftarrow$ makeNewOffspringPop($P_t$)
    $R_t \leftarrow P_t + Q_t$
    while $|R_t| > |P_t|$ do
        $I \leftarrow$ getWeakestIndividual($R_t$)
        $R_t \leftarrow R_t - I$
    end while
    $P_t \leftarrow R_t$
    $t \leftarrow t + 1$
end while
```

The procedure *getWeakestIndividual* of determining the weakest individual in Algorithm 1 is defined as follows:

1. Systematically remove one individual
2. Compute the weighted sum score according to Equation 1
3. Add the individual back to the population
4. Repeat from step 1. until all individuals have been tried once and the worst individual can be determined.

The weighted sum score of a given population P is calculated as follows:

$$WSM_{score}(P) = -W_{fit} \cdot F_{fit}(P) + W_{div} \cdot F_{div}(P) \tag{1}$$

Note the negation of W_{fit} in Equation 1 as we want to maximise diversity but also minimise fitness at the same time.

The population fitness F_{fit} is computed by simply taking the average of the fitness of all M individuals of the current population P:

$$F_{fit}(P) = \frac{1}{M}\sum_{i=1}^{M} F(x) \tag{2}$$

An effective and simple measure of distance between two sequences is the Hamming-distance. For two sequences $A = \{aa_i\}$ and $A' = \{aa'_i\}$ where $1 \leq i \leq N$, the normalised Hamming distance between them is defined as:

$$d_{Hamm}(A, A') = \frac{1}{N}\sum_{i=1}^{N} d_i \quad where \quad d_i = \begin{cases} 0 & \text{if } aa_i = aa'_i \\ 1 & \text{if otherwise} \end{cases} \tag{3}$$

The population diversity F_{div} is computed by taking the average Hamming distance of each M individuals to the remaining $M-1$ individuals of the population P:

$$F_{div}(P) = \frac{1}{M \cdot (M-1)}\sum_{i=1}^{M}\sum_{j=1}^{M} d_{Hamm}(A_i, A_j), \quad \forall i \neq j \tag{4}$$

5 Experimental Results

To study the performance of the PBGA with respect to fitness and diversity convergence, a number of experiments have been conducted to compare it against different Genetic Algorithms, i.e., the generational (gGA), the synchronous cellular (scGA) and the steady-state (ssGA). The PBGA was tested with the following six different weight ratio settings:

$$W_{(fit,div)} = \{(1.0, 0.0), (0.9, 0.1), (0.8, 0.2), (0.7, 0.3), (0.5, 0.5), (0.3, 0.7)\}.$$

Table 2 summarises the settings and parameters used to conduct the experiments.

Figure 1a illustrates the convergence of fitness for the best performing PBGA setting in comparison with the gGA, scGA and the ssGA. The gGA performs the worst and the PBGA with a weight setting of (0.9, 0.1) surpasses the ssGA and achieves better final fitness results than all of the other GAs. Figure 1b illustrates the diversity convergence for the same algorithms. It is noted that the PBGA achieves a higher diversity than the scGA and ssGA while at the same time having better fitness results. Similar graphs are obtained for the NK-IFP-2 model and are hence not shown here.

Table 3 summarises average fitness and diversity for all the algorithms tested highlighting best and worst algorithm results in light and dark grey respectively. With a weight setting of (0.9, 0.1) the PBGA achieves the best fitness for both

Table 2: Experimental settings.

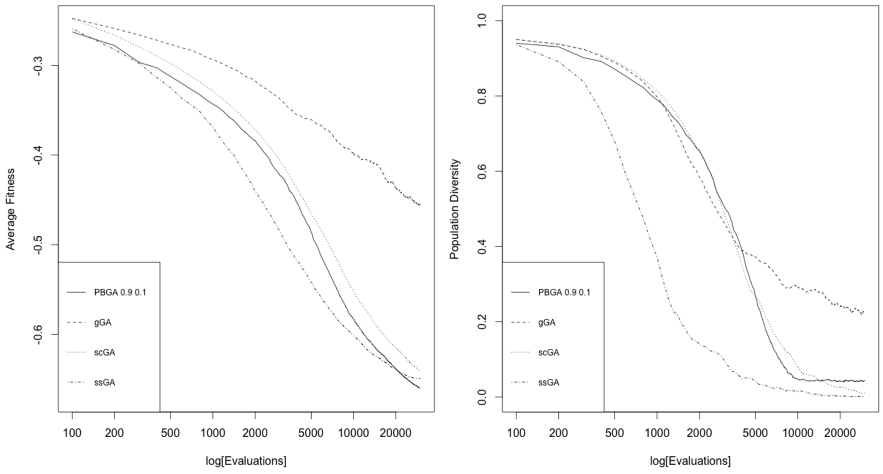
Setting	**Value**
GAs	gGA, scGA, ssGA and PBGA
Population size	100
Termination condition	30000 function evaluations
Number of independent runs	30
Selection	Binary tournament (BT)
Neighbourhood	C9 in scGA
Crossover operator	SPX, $p_c = 0.9$
Mutation operator	Uniform, $p_m = \frac{1}{N}$
Elitism	2 individuals (for gGA)

benchmark models with -0.662 for the best value and -0.660 on average for model 1 and with -0.632 for the best value and -0.631 on average for model 2. It is interesting to note that the PBGA with a weight setting of (0.5, 0.5) achieves better results than the gGA in terms of fitness as well as diversity for both models with -0.574 vs. -0.559 for the best fitness value and -0.511 vs. -0.456 on average for model 1 and with -0.550 vs. -0.545 for the best fitness value and -0.485 vs. -0.429 on average for model 2.

In order to provide statistical confidence, the Wilcoxon test indicator was applied with a 5% significance level. With a weight setting of (0.9, 0.1), the PBGA clearly outperforms the gGA and the scGA with statistical confidence for the average fitness with values -0.662 vs. -0.559 and -0.662 vs. -0.644 respectively for model 1 and with values -0.632 vs. -0.545 and -0.632 vs. -0.621 respectively for model 2, whereas in comparison with the ssGA the PBGA does not achieve as quick good results as the ssGA, but surpasses the ssGA in the end and achieves better average fitness values of -0.662 vs. -0.650 respectively for model 1 and with values -0.632 vs. -0.628 respectively for model 2. However, as seen in Figure 1a for model 1, the final slope is steeper than the ssGA, indicating better performance. The steeper final slope can be explained by the constantly high diversity as seen in Figure 1b for model 1, which allows for continued exploration while the other GAs suffer from premature convergence.

Table 3: Final values in terms of fitness and diversity averaged over 30 independent runs for the two NK benchmark models.

	Model 1				**Model 2**			
	Fitness		**Diversity**		**Fitness**		**Diversity**	
Algorithm	**Best**	**Average**	**Best**	**Average**	**Best**	**Average**	**Best**	**Average**
$\text{PBGA}_{1.0\ 0.0}$	-0.649	-0.648 $\pm 0.37E{-3}$	0.005	0.002 $\pm 1.78E{-3}$	-0.628	-0.628 $\pm 0.27E{-3}$	0.004	0.001 $\pm 1.56E{-3}$
$\text{PBGA}_{0.9\ 0.1}$	-0.662	-0.660 $\pm 1.07E{-3}$	0.041	0.043 $\pm 0.84E{-3}$	-0.632	-0.631 $\pm 0.76E{-3}$	0.031	0.038 $\pm 3.25E{-3}$
$\text{PBGA}_{0.8\ 0.2}$	-0.652	-0.627 $\pm 12.3E{-3}$	0.250	0.337 $\pm 43.8E{-3}$	-0.621	-0.594 $\pm 13.4E{-3}$	0.310	0.406 $\pm 48.1E{-3}$
$\text{PBGA}_{0.7\ 0.3}$	-0.629	-0.582 $\pm 23.3E{-3}$	0.508	0.612 $\pm 52.1E{-3}$	-0.602	-0.557 $\pm 22.5E{-3}$	0.542	0.639 $\pm 48.7E{-3}$
$\text{PBGA}_{0.5\ 0.5}$	-0.574	-0.511 $\pm 31.4E{-3}$	0.774	0.833 $\pm 31.4E{-3}$	-0.550	-0.485 $\pm 32.3E{-3}$	0.787	0.846 $\pm 29.6E{-3}$
$\text{PBGA}_{0.3\ 0.7}$	-0.527	-0.458 $\pm 34.5E{-3}$	0.880	0.909 $\pm 14.6E{-3}$	-0.503	-0.440 $\pm 31.7E{-3}$	0.888	0.913 $\pm 12.4E{-3}$
gGA	-0.559	-0.456 $\pm 51.4E{-3}$	0.145	0.227 $\pm 40.9E{-3}$	-0.545	-0.429 $\pm 58.0E{-3}$	0.138	0.221 $\pm 41.1E{-3}$
scGA	-0.644	-0.641 $\pm 1.54E{-3}$	0.017	0.010 $\pm 3.36E{-3}$	-0.621	-0.619 $\pm 1.20E{-3}$	0.013	0.009 $\pm 2.18E{-3}$
ssGA	-0.650	-0.645 $\pm 0.18E{-3}$	0.005	0.001 $\pm 1.97E{-3}$	-0.628	-0.628 $\pm 0.14E{-3}$	0.001	0.001 $\pm 0.42E{-3}$



(a) Fitness convergence PBGA vs. GAs (b) Diversity convergence PBGA vs. GAs

Fig. 1: NK benchmark model NK-IFP-1 average fitness and diversity convergence.

6 Conclusion

In this paper a novel Preference-Based Genetic Algorithm (PBGA) was presented in combination with a weighted sum model, which allows to shift focus arbitrarily between diversity and fitness with a direct effect on the population as a whole without relying on secondary effects from added mechanisms or operators. The PBGA was evaluated on NK benchmark models and compared to state-of-the-art GAs. Final results were found comparable or better than the other GAs on average, while the diversity of found sequences remains higher at the same time. The best results were achieved using a weight setting of (0.9, 0.1) where 0.9 represents 90% of fitness and 0.1 represents 10% of diversity. In addition, the PBGA showed a better convergence, which promises even better solutions, given an evaluation budget beyond the computational limitations set in this work. Future work will focus on the development of a more advanced preference evaluation model using Fuzzy logic while adding more preferences such as crowding or elitism, and making the selection of preferences adaptive.

References

1. E. Alba and B. Dorronsoro. The exploration/exploitation tradeoff in dynamic cellular genetic algorithms. *Evolutionary Computation, IEEE Transactions on*, 9(2):126–142, 2005.
2. K. A. De Jong. Analysis of the behavior of a class of genetic adaptive systems. 1975.

3. D. E. Goldberg and J. Richardson. Genetic algorithms with sharing for multimodal function optimization. In *Genetic algorithms and their applications: Proceedings of the Second International Conference on Genetic Algorithms*, pages 41–49. Hillsdale, NJ: Lawrence Erlbaum, 1987.
4. J. L. Jimenez Laredo, S. S. Nielsen, G. Danoy, P. Bouvry, et al. Cooperative selection: Improving tournament selection via altruism. In *The 14th European Conference on Evolutionary Computation in Combinatorial Optimisation*, 2014.
5. A. López-Jaimes and C. A. C. Coello. Including preferences into a multiobjective evolutionary algorithm to deal with many-objective engineering optimization problems. *Information Sciences*, 277:1–20, 2014.
6. S. S. Nielsen, G. Danoy, P. Bouvry, and E.-G. Talbi. Nk landscape instances mimicking the protein inverse folding problem towards future benchmarks. In *Proceedings of the Companion Publication of the 2015 on Genetic and Evolutionary Computation Conference*, GECCO Companion '15, pages 915–921, New York, NY, USA, 2015. ACM.
7. H. Shimodaira. Dcga: A diversity control oriented genetic algorithm. In *Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on*, pages 367–374. IEEE, 1997.
8. L. Thiele, K. Miettinen, P. J. Korhonen, and J. Molina. A preference-based evolutionary algorithm for multi-objective optimization. *Evolutionary Computation*, 17(3):411–436, 2009.
9. E. Zitzler and S. Künzli. Indicator-based selection in multiobjective search. In *Parallel Problem Solving from Nature-PPSN VIII*, pages 832–842. Springer, 2004.

Stable models in normal residuated logic programs

M. Eugenia Cornejo[1], David Lobo[2], and Jesús Medina[2]

[1]Department of Statistic and O.R., University of Cádiz. Spain
mariaeugenia.cornejo@uca.es
[2]Department of Mathematics, University of Cádiz. Spain
{david.lobo,jesus.medina}@uca.es

Abstract. The existence of stable models for a normal residuated logic program defined on $[0, 1]$ and the uniqueness of these models in the particular case of the product t-norm, its residuated implication, and the standard negation have been recently studied by Madrid and Ojeda-Aciego [10]. In this paper, we introduce results which generalize the existence of stable models for normal residuated logic programs defined on any convex compact set of an euclidean space. In addition, we show which conditions are required in order to guarantee the uniqueness of a stable model for a normal residuated logic program defined on $\mathcal{C}([0, 1])$.

Key words: negation; normal residuated logic program; stable model.

1 Introduction

Searching for conditions guarantying the existence and uniqueness of fuzzy stable models in normal residuated logic programming has received a strong attention since the definition of this kind of programs [1].

However, the existence of stable models cannot be guaranteed for an arbitrary normal residuated logic program [2]. This is due to the fact that fuzzy framework includes two different dimmensions: the syntactic structure of the normal program (the syntaxis) and the choice of suitables connectives in the underlying lattice, the semantics of the program.

As the connectives are fixed in classical logic programming, we can only establish the syntactic conditions of the program. Nevertheless, we can choose many operators to use them as connectives in normal residuated logic programs, and this implies that semantics plays a crucial role in this framework.

Until now, only a few sufficient conditions have been found to ensure the existence of fuzzy stable models in some approaches. In [3], it has been proven that every normal logic program has stable models in the 3-valued Kleene logic. Furthermore, by [4–8], we know that every normal residuated logic program has stable models if the underlying residuated lattice has an appropriate bilattice structure [9]. Recently, it has been shown in [10] conditions to ensure the existence and unicity of stable models for a normal residuated logic program defined on $[0, 1]$.

In this paper, we will generalize the result of existence of stable models for programs defined on any convex compact set of an euclidean space. Moreover, we will introduce the conditions which guarantee the uniqueness of stable models for normal residuated logic programs defined on $\mathcal{C}([0,1])$.

2 Preliminaries

In this section, we will recall the main definitions and results which will be used in the paper. Firstly, we introduce the definition of residuated lattice.

Definition 1. *A* residuated lattice *is a tuple* $(L, \leq, *, \leftarrow)$ *such that:*

(1) $(L, \leq)$ *is a complete bounded lattice with* $\top$ *and* $\bot$ *the greatest and the least elements, respectively;*
(2) $(\leftarrow, *)$ *is an adjoint pair in* $(L, \leq)$*, that is, the equivalence:*

$$z \leq (x \leftarrow y) \quad \textit{if and only if} \quad y * z \leq x$$

holds, for all $x, y, z \in L$.
(3) $(L, *, \top)$ *is a commutative monoid.*

Note that, the adjoint pair is uniquely determined by the chosen operator $*$. Specifically, fixed a left-continuous operator $*$, its adjoint implication is defined as follows:

$$x \leftarrow y = \sup\{z \in L : y * z \leq x\}$$

Now, we will consider a residuated lattice enriched with a negation operator. A negation operator is any decreasing mapping $n : L \to L$ satisfying $n(\bot) = \top$ and $n(\top) = \bot$. The negation will model the notion of default negation.

Definition 2. *Given a residuated lattice with negation* $(L, \leq, *, \leftarrow, \neg)$*, a* normal residuated logic program $\mathbb{P}$ *is a finite set of weighted rules of the form:*

$$\langle p \leftarrow p_1 * \cdots * p_m * \neg p_{m+1} * \cdots * \neg p_n; \quad \vartheta \rangle$$

where ϑ *is an element of* L *and* $p, p_1, \ldots, p_n$ *are propositional symbols such that* $p_i \neq p_j$*, for all* $i, j \in \{1, \ldots, n\}$.

As usual, we denote the rules as $\langle p \leftarrow \mathcal{B}; \vartheta \rangle$, where p is the *head* of the rule, $\mathcal{B}$ its *body* and ϑ its *weight*. A *fact* is a rule where no propositional symbols appear in the body.

The set of propositional symbols appearing in P is denoted by $\Pi_{\mathbb{P}}$.

Definition 3. *A fuzzy* L-interpretation *is a mapping* $I : \Pi_{\mathbb{P}} \to L$ *which assigns a truth value to every propositional symbol appearing in* P*. We say that:*

(1) I *satisfies a rule* $\langle p \leftarrow \mathcal{B}; \vartheta \rangle$ *if and only if* $\vartheta \leq I(p \leftarrow \mathcal{B})$.
(2) I *is a model of* $\mathbb{P}$ *if it satisfies all rules in* $\mathbb{P}$.

The set of all L-interpretations will be denoted as $I_{\mathfrak{U}}$, where $\mathfrak{U}$ is the residuated algebra in which the lattice is defined. An ordering relation $\sqsubseteq$ can be defined in $I_{\mathfrak{U}}$ as follows: Given I and J two L-interpretations, $I \sqsubseteq J$ if and only if $I(p) \leq J(p)$, for all $p \in \Pi_{\mathbb{P}}$.

Given a finite normal residuated logic program $\mathbb{P}$ defined on L, the set of L-interpretations with the new ordering relation verifies some properties of the underlying lattice. Specifically, it inherits the properties of the cartesian product of several copies of the lattice. Indeed, each L-interpretation can be seen as an element of L^n, where n is the cardinal of $\Pi_{\mathbb{P}}$.

Theorem 1. *If $\langle L, \leq \rangle$ is a complete lattice, then $\langle I_{\mathfrak{U}}, \sqsubseteq \rangle$ is a complete lattice.*

2.1 Immediate consequence operator and stable models

A generalization of the immediate consequence operator for normal residuated logic programs is given in the next definition.

Definition 4. *Let $\mathbb{P}$ be a normal residuated logic program. The* immediate consequence operator *is the mapping $T_{\mathbb{P}} : I_{\mathfrak{U}} \to I_{\mathfrak{U}}$ defined as*

$$T_{\mathbb{P}}(I)(p) = \sup\{I(\mathcal{B}) * \vartheta : \langle p \leftarrow \mathcal{B}; \vartheta \rangle \in \mathbb{P}\}$$

where $p \in \Pi_{\mathbb{P}}$.

If $\mathbb{P}$ is a positive program (without any negation), then $T_{\mathbb{P}}$ is a monotonic operator and we can characterize the models of the residuated program by the post-fix points of $T_{\mathbb{P}}$.

Proposition 1. *Let $\mathbb{P}$ be a positive residuated logic program. Then M is a model of $\mathbb{P}$ if and only if $T_{\mathbb{P}}(M) \leq M$.*

Knaster-Tarski's fix point theorem ensures that the operator $T_{\mathbb{P}}$ has a least fix point. In addition, by the proposition above, this least fix point is actually the least model of $\mathbb{P}$. This fact leads us to define the least model semantics in positive residuated logic programs.

The main difference with respect to the case of normal residuated logic programs is that $T_{\mathbb{P}}$ is not necessarily monotonic. Therefore, we cannot guarantee the existence of the least model and we need another notion to define the semantics for a normal residuated logic program. A new mathematical object which generalizes the least model semantics to normal residuated logic programs is required. This object is the stable model of a program, which was defined in [11].

Let $\mathbb{P}$ be a normal residuated logic program and I a fuzzy L-interpretation. First of all, we will build a positive residuated program $\mathbb{P}_I$ by substituting each rule in $\mathbb{P}$ such as

$$\langle p \leftarrow p_1 * \cdots * p_m * \neg p_{m+1} * \cdots \neg p_n; \quad \vartheta \rangle$$

by the rule

$$\langle p \leftarrow p_1 * \cdots * p_m; \quad \neg I(p_{m+1}) * \cdots * \neg I(p_n) * \vartheta \rangle$$

Observe that, we can apply to $\mathbb{P}_I$ the known results to positive residuated program.

Definition 5. *The program $\mathbb{P}_I$ is called the* reduct *of $\mathbb{P}$ with respect to the interpretation I.*

Thanks to the notion of reduct we can define a stable model of a program.

Definition 6. *Let $\mathbb{P}$ be a normal residuated logic program and let I be a fuzzy L-interpretation. I is said to be a stable model of $\mathbb{P}$ if and only if I is a minimal model of $\mathbb{P}_I$.*

An important feature of stable models, which is also verified in our framework, is that a stable model is always a minimal fix point of $T_{\mathbb{P}}$.

Proposition 2. *Any stable model of $\mathbb{P}$ is a minimal fix point of $T_{\mathbb{P}}$.*

It is worth noting that the counterpart of Proposition 2 is not satisfied, in general, because the $T_{\mathbb{P}}$ operator is not necessarily monotonic.

3 On the existence and unicity of stable models

Our goal is to extend the obtained results by Madrid and Ojeda-Aciego about the existence and the unicity of stable models for normal residuated logic programs [10]. With this purpose, we need to consider an extension of Brouwer's fix point theorem.

Theorem 2. *Let X an euclidean space and let K be a a convex compact set not empty. Every continuous mapping $f\colon K \to K$ has a fix point.*

Note that, the set of all L-interpretations of a normal residuated logic program defined on a lattice with convex (compact, respectively) support is a convex (compact, respectively) set. This fact leads us to present the following result.

Proposition 3. *Let $\mathbb{P}$ be a normal residuated logic program defined on a lattice $(K, \leq, *, \leftarrow, \neg)$ where K is a convex (compact, respectively) set in an euclidean space X. Then the set of L-interpretations of $\mathbb{P}$ is a convex (compact, respectively) set in the set if mappings defined on X.*

Applying Theorem 2 to the operator R defined by $R(I) = T_{\mathbb{P}_I}$, we obtain that $T_{\mathbb{P}_I}$ have a fix point. As $\mathbb{P}_I$ is a positive residuated logic program, we obtain that this fix point is actually the minimal model of $\mathbb{P}_I$ and then it is a stable model of $\mathbb{P}$.

The continuity of the connectives $*$ and $\neg$ plays a key role in order to apply the Theorem 2 to the operator R.

Theorem 3. *Let $(K, \leq, *, \leftarrow, \neg)$ be a residuated lattice where K is a convex compact non-empty set in an euclidean space. If $*$ and $\neg$ are continuous operators, then every finite normal program $\mathbb{P}$ defined on this lattice has at least a stable model.*

Finally, we present a result which ensure the uniqueness of the stable models for a normal program defined with the product adjoint pair and the standard negation on the set of subintervals of $[0, 1]$, that is, $\mathcal{C}([0, 1])$. This fact is interesting because not only one truth value can be assigned to each propositional symbol in $\mathbb{P}$, but we can assign a minimal truth value and a maximal truth value for the propositional symbol.

Theorem 4. *Let $\mathbb{P}$ be a normal residuated logic program defined on $\mathcal{C}([0, 1])$, and let us write for each propositional symbol p in $\mathbb{P}$, $\vartheta_p = \max\{\vartheta_j : \langle p \leftarrow \beta\ ; \vartheta_j\rangle \in \mathbb{P}\}$. Then, if for every rule $\langle p \leftarrow q_1 * \cdots * q_h * \neg q_{h+1} * \cdots * \neg q_k\ ; \vartheta\rangle \in \mathbb{P}$, the inequality below holds*

$$\left(\sum_{j=1}^{h} \vartheta_{q_1} \cdot \ldots \cdot \vartheta_{q_{j-1}} \cdot \vartheta_{q_{j+1}} \cdot \ldots \cdot \vartheta_{q_h} \cdot \vartheta\right) + (k - h)(\vartheta_{q_1} \cdot \ldots \cdot \vartheta_{q_h} \cdot \vartheta) < (1, 1)$$

then there exists only one stable model of $\mathbb{P}$.

4 Conclusions

We have shown results which guarantee the existence of stable models for normal programs defined on a convex compact set, and which guarantee the uniqueness of stable models for normal programs defined on $\mathcal{C}([0, 1])$.

References

1. C., Damásio and L., Moniz Pereira. Monotonic and residuated logic programs. Sixth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU01), pp. 748–759. Springer Verlag (2001).
2. N., Madrid and M., Ojeda-Aciego. On coherence and consistence in fuzzy answer set semantics for residuated logic programs. Lect. Notes in Computer Science 5571, pp. 60–67, (2009).
3. T., Przymusinski. Well-founded semantics coincides with three-valued stable semantics. Fundamenta Informaticae 13, pp. 445–463, (1990).
4. M., Fitting. The family of stable models. The Journal of Logic Programming 17(2-4), pp. 197–225, (1993).
5. Y., Loyer and U., Straccia. Epistemic foundation of stable model semantics. Journal of Theory and Practice of Logic Programming 6, pp. 355–393, (2006).
6. U., Straccia. Query answering in normal logic programs under uncertainty. Lect. Notes in Computer Science 3571, pp. 687–700, (2005).
7. U., Straccia. Query answering under the any-world assumption for normal logic programs. Lect. Notes in Computer Science 3571, pp. 687–700, (2006).

8. U., Straccia. A top-down query answering procedure for normal logic programs under the any-world assumption. Proc. of the 10th Intl Conf on Principles of Knowledge Representation, pp. 329–339, (2006).
9. M. L., Ginsberg. Multivalued logics: a uniform approach to reasoning in artificial intelligence. Computational Intelligence 4, pp. 265–316, (1988).
10. N., Madrid and M., Ojeda-Aciego. On the existence and unicity of stable models in normal residuated logic programs. International Journal of Computer, Mathematics (2012).
11. N., Madrid and M., Ojeda-Aciego. Towards a fuzzy answer set semantics for residuated logic programs. Web Intelligence/IAT Workshops, pp. 260–264, (2008).

A Study of Multiple Sequence Alignment With Multi-Objective Metaheuristics

Antonio J. Nebro[1], Cristian Zambrano-Vega[2] Juan J. Durillo[3], and José Francisco Aldana Montes[1]

[1] Edificio de Investigación Ada Byron, University of Málaga, Spain,
[2] Facultad de Ciencias de la Ingeniería, Universidad Técnica Estatal de Quevedo,
Quevedo - Los Ríos - Ecuador
[3] Distributed and Parallel Systems Group, University of Innsbruck, Austria

Abstract. Multiple sequence alignment (MSA) is a problem from the bioinformatics domain consisting in finding the best possible alignment for a set of three or more sequences. Different scores have been proposed to assess the quality of MSA solutions, so the problem can be formulated as a multi-objective optimization problem. In this paper we carry out a performance study involving five multi-objective metaheuristics which are representative of the state-of-the-art. The results when solving a number of instance problems reveals that the classical NSGA-II and SPEA2 algorithms can outperform more modern techniques.

Keywords: Multiple sequence alignment, multi-objective optimization, metaheuristics, performance comparison

1 Introduction

The alignment of multiple DNA, RNA and protein sequences (MSA) is a common task in Bioinformatics [1]. The aim of MSA is comparing different sequences in order to extract their shared information and their significant differences. The alignment of pair of sequences can be achieved by using dynamic programming techniques, but these strategies cannot be applied when dealing with three or more sequences because the search space grows exponentially with the number of sequences and it is also dependent on the sequence lengths. These reasons have lead to the use metaheuristics to deal with MSA problems [2].

An additional issue in MSA is that there exist different methods to measure the accuracy on an alignment, so the problem can be formulated as a multi-objective optimization problem [3][4]. The motivation of our work is that these studies rely on the use of the NSGA-II algorithm [5], so we are interested in determining whether other algorithms could be more adequate for solving MSA problems. In this paper, we elaborate our first approximation to this issue. Our main contribution is the comparison of a number of multi-objective metaheuristics: NSGA-II, SPEA2 [6], AbYSS [7], MOCell [8], and SMS-EMOA [9]. All these algorithms but NSGA-II are applied the first time to MSA to the best of our knowledge.

The rest of the paper is organized as follows. Section 2 includes a review of related work. The problem is described in Section 3. The experimentation details and an analysis of the obtained results are presented in Sections 4 and 5. Finally, the conclusions and lines of future work are commented in Section 6.

2 Related Work

In this section we briefly review some multi-objective approaches published in the literature to solve the MSA problem using multi-objective optimization techniques.

Ortuño *et al.* implemented a NSGA-II based multi-objective evolutionary algorithm to align multiple sequences and applied it to optimizing three objectives: STRIKE score, non-gaps percentage and totally conserved columns [3]. Soto and Becerra proposed a multi-objective evolutionary algorithm, also inspired in NSGA-II, to optimize pre-aligned sequences in [10]. They used two objectives functions to compare the quality of the MSA: the entropy and the MetAl metrics. A multi-objective genetic algorithm based in NSGA-II (MSAGMOGA) is described in [4], where three objectives are considered: similarity, affine gap penalty and support.

The first two works take the approach of pre-computing alignments with existing tools (Muscle, ClustalW, Mafft, T-Coffee, etc.), in such a way that the initial populations contain aligned solutions. We use also this idea in this work.

It is worth nothing that these three papers consider different objectives, so there is no a consensus about how assess the quality of the alignments. This makes also makes difficult to compare new proposals against them.

3 Problem Description

Given a finite alphabet Σ and a set $S = (s_1, s_2, ..., s_n)$ of n sequences of varying length, an alignment is a matrix where all the symbols of the sequences appear in the same order and a special symbol or gap (typically represented with the character '-') can be inserted potentially at any position in such a way that all the sequences have the same length.

An example of alignment is shown below, representing four sequences with two aligned columns (marked with an asterisk).

```
APPSVFAEVPQ-AQPV
AKRS-V-E-PFR-IKM
-LISKRA-YP--I---
-SASTIGVEPC-RA-P
   *     *
```

The MSA problem consists then in inserting gaps in the proper places in order to maximize some scores. For example, in [3] two of the considered objectives are to maximize the non-gaps percentage and the percentage of completely aligned columns. These are very intuitive goals, but they are not fully contradictory from a multi-objective point of view: if after manipulating the sequences a column is full of gaps then it can be removed, thus improving the number of non-gaps, but this does not implies a worsening in the percentage of aligned columns.

In this work we have select two objectives: the aforementioned percentage of aligned columns and the sum of pairs (SOP), which is computed by adding all the scores of the pairwise comparisons between each symbol in each column of the alignment. A scoring matrix is need to calculate the SOP; we have used the PAM250 matrix (with a gap penalty of -8).

4 Experimentation

In this section, we briefly describe the algorithms we have selected, the chosen benchmark, and experimentation methodology.

Our study includes five multi-objective metaheuristics. NSGA-II [5] and SPEA2 [6] are classical evolutionary algorithms which have been widely used since they were proposed. SMS-EMOA [9] and MOCell [8] are also evolutionary algorithms and they are representative of indicator-based and cellular techniques, respectively. The last method is AbYSS [7], a scatter search algorithm. All the metaheuristics have been implemented in the jMetal framework [11].

We have considered a set of parameter settings that have been adopted in other studies. This way, all the algorithms runs until 25000 function evaluations have computed, the population sizes have a size of 100 in the evolutionary algorithms (20 in AbYSS), and MOCell and AbYSS has an archive size of 100. All the metaheuristics include the same genetic operators: single-point crossover (applied with a probability of 1.0) and a multiple mutation operator which randomly selects one out of three mutations: one gap insertion (a gap is randomly inserted), one gap shifting (a random gap is selected and it is shifted with the symbol on the right or on the left), and gaps merging (a number of gaps are joined to appear consecutively in the sequence). This multiple mutation operator is applied with a probability of $1.0/L$, where L is the number of sequences. The encoding used to represent the sequences consists of lists of characters.

As commented before, each problem has been previously aligned with a number of tools, namely Clustal Omega, T-Coffee, Mafft, and Muscle. The obtained alignments are included in the initial population of all the algorithms, and they are also used to create new solutions. The process consists on choosing an alignment and inserting a number of gaps (from 1 to 5) at random positions.

We have chosen five problems from the BAliBASE 3.0 library [12]. Concretely, we have taken five instances of the RV11 reference set, which are referred as to BB11001 to BB11005. They range from 4 to 14 sequences.

The experimentation methodology is described next. 20 independent runs have been carried of out of each combination algorithm-problem, and the Hypervolume quality indicator [13] has been computed to all the yielded Pareto front approximations. We report the obtained median and interquartile range (IQR) values. To check the significance of the differences between the algorithms we have applied the unpaired Wilcoxon rank-sum test with a confidence level

Table 1. Hypervolume quality indicator values. Median and IQR

	NSGAII	SPEA2	ABYSS	SMS-EMOA	MOCell
BB11001	$0.00e+00_{0.0e+00}$	$0.00e+00_{1.5e-01}$	$1.46e-03_{5.7e-02}$	$0.00e+00_{0.0e+00}$	$0.00e+00_{0.0e+00}$
BB11002	$4.04e-01_{2.4e-01}$	$4.97e-01_{1.2e-01}$	$1.46e-01_{1.2e-01}$	$4.99e-01_{1.2e-01}$	$1.66e-01_{2.5e-02}$
BB11003	$2.28e-01_{1.6e-01}$	$2.63e-01_{2.2e-01}$	$0.00e+00_{0.0e+00}$	$1.91e-01_{1.4e-01}$	$0.00e+00_{0.0e+00}$
BB11004	$2.93e-01_{1.8e-01}$	$2.80e-01_{9.3e-02}$	$0.00e+00_{0.0e+00}$	$2.03e-01_{1.5e-01}$	$4.83e-02_{7.0e-02}$
BB11005	$4.44e-01_{1.3e-02}$	$4.08e-01_{1.5e-02}$	$0.00e+00_{0.0e+00}$	$4.07e-01_{2.5e-02}$	$3.90e-01_{3.5e-02}$

Table 2. Results of the Wilcoxon rank-sum test. Each symbol in the cells represents the five considered ploblems. The ▲ symbol indicates that the algorithm in the row is significantly better than the algorithm in the column, a ▽ means the opposite, and a '-' states that the differences are non-significant.

	SPEA2	ABYSS	SMSEMOA	MOCell
NSGAII	– – – – ▲	▽ ▲ ▲ ▲ ▲	– – – – ▲	– ▲ ▲ ▲ ▲
SPEA2		– ▲ ▲ ▲ ▲	– – – – –	▲ ▲ ▲ ▲ ▲
ABYSS			– ▽ ▽ ▽ ▽	▲ – – ▽ ▽
SMSEMOA				▲ ▲ ▲ ▲ ▲

of 95% (i.e., significance level of 5% or p-value under 0.05), meaning that the differences are unlikely to have occurred by chance with a probability of 95%.

As the true Pareto fronts of the solved problems are unknown, we have built a reference front for each problem by joining all the fronts obtained by all the algorithms in all the independent runs and deleting the dominated solutions.

5 Results

The obtained results are included in Table 1, where the cells with dark and light gray background colors indicate respectively the best and second best values. We can observe than some cells contain a value of 0; this means that the obtained fronts are beyond the limits of the reference Pareto front, so the Hypervolume is not computed in those situations.

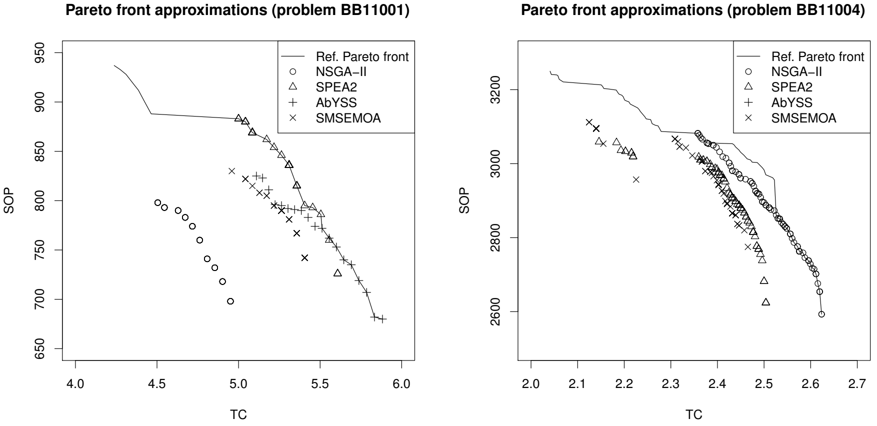
The values included in the table reveals that NSGA-II and SPEA2 have the best overall performance, outperforming the other algorithms. Both NSGA-II and SPEA2 are generational evolutionary algorithms, while SMS-EMOA and MOCell follow a steady-state scheme, so a first conclusion could be that the generational selection scheme can have a positive influence in searching for optimal alignments. However, the results of the Wilcoxon rank-sum test (see Table 2) show that most of the differences in the pairwise comparison between NSGA-II, SPEA2, and SMS-EMOA are not significant.

To illustrate the Pareto front approximations that are found by the compared algorithms (we have excluded MOCell), we include in Figure 1 the fronts having the best Hypervolume value for problems BB11001 and BB11004.

6 Conclusions and Future Work

We have presented a study of solving MSA problems with a number of multi-objective metaheuristics. Our main motivation has been that multi-objective

Fig. 1. Pareto front approximations having the best Hypervolume obtained by NSGA-II, SPEA2, SMS-EMOA, and AbYSS for problems BB11001 and BB11004.



optimization approaches for MSA are scarce in the literature, and practically all of them relies on the use of NSGA-II. So, we have selected five multi-objective metaheuristics and we have compared them again a benchmark of five problems.

The conclusion is that, according the parameter settings we have used and the chosen benchmark, the classic NSGA-II and SPEA2 algorithms outperform more recent proposals according to the Hypervolume quality indicator.

Our work is a first step in the open issue of MSA with multi-objective metaheuristics. First, we have used standard settings established in other studies (mainly on continuous optimization), so a parameter sensitivity study is needed to find more effective algorithm configurations to deal with MSA. Second, the benchmark must be augmented in a significant way to draw firmer conclusions. Finally, an analysis of the objectives to optimize must be carried out to decide which scores are really important; if more than four or five would be of interest, the MSA would become a many-objective problem, so a new set of metaheuristics should be needed.

Acknowledgments. This work is partially funded by Secretaría Nacional de Educación Superior Ciencia y Tecnología SENESCYT from Ecuador and by Grants TIN2014-58304-R (Ministerio de Ciencia e Innovación) and P11-TIC-7529 and P12-TIC-1519 (Plan Andaluz de Investigación, Desarrollo e Innovación).

References

1. Pei, J.: Multiple protein sequence alignment. Current Opinion in Structural Biology **18**(3) (2008) 382 – 386 Nucleic acids / Sequences and topology.

2. da Silva, F., Pérez, J.S., Pulido, J.G., Rodríguez, M.V.: Alineagaa genetic algorithm with local search optimization formultiple sequence alignment. Applied Intelligence **32**(2) (2010) 164–172
3. Ortuno, F., Valenzuela, O., Rojas, F., Pomares, H., Florido, J., Urquiza, J., Rojas, I.: Optimizing multiple sequence alignments using a genetic algorithm based on three objectives: structural information, non-gaps percentage and totally conserved columns. Bioinformatics (Oxford, England) **29**(17) (September 2013) 2112–21
4. Kaya, M., Sarhan, A., Abdullah, R.: Multiple sequence alignment with affine gap by using multi-objective genetic algorithm. Computer methods and programs in biomedicine **114**(1) (April 2014) 38–49
5. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation **6**(2) (2002) 182–197
6. Zitzler, E., Laumanns, M., Thiele, L.: SPEA2: Improving the strength pareto evolutionary algorithm. In Giannakoglou, K., Tsahalis, D., Periaux, J., Papailou, P., Fogarty, T., eds.: EUROGEN 2001. Evolutionary Methods for Design, Optimization and Control with Applications to Industrial Problems, Athens, Greece (2002) 95–100
7. Nebro, A., Luna, F., Alba, E., Dorronsoro, B., Durillo, J., Beham, A.: AbYSS: Adapting Scatter Search to Multiobjective Optimization. IEEE Transactions on Evolutionary Computation **12**(4) (August 2008)
8. Nebro, A., Durillo, J., Luna, F., Dorronsoro, B., Alba, E.: Design issues in a multiobjective cellular genetic algorithm. In Obayashi, S., Deb, K., Poloni, C., Hiroyasu, T., Murata, T., eds.: Evolutionary Multi-Criterion Optimization. 4th International Conference, EMO 2007. Volume 4403 of Lecture Notes in Computer Science., Springer (2007) 126–140
9. Emmerich, M., Beume, N., Naujoks, B.: An emo algorithm using the hypervolume measure as selection criterion. In Coello, C., Hernández, A., Zitler, E., eds.: Third International Conference on Evolutionary MultiCriterion Optimization, EMO 2005. Volume 3410 of LNCS., Springer (2005) 62–76
10. Soto, W., Becerra, D.: A multi-objective evolutionary algorithm for improving multiple sequence alignments. In Campos, S., ed.: Advances in Bioinformatics and Computational Biology. Volume 8826 of Lecture Notes in Computer Science. Springer International Publishing (2014) 73–82
11. Durillo, J., Nebro, A.: jmetal: A java framework for multi-objective optimization. Advances in Engineering Software **42**(10) (2011) 760 – 771
12. Thompson, J., Koehl, P., Poch, O.: Balibase 3.0: latest developments of the multiple sequence alignment benchmark. Proteins **61** (2005) 127–136
13. Zitzler, E., Thiele, L.: Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. IEEE Transactions on Evolutionary Computation **3**(4) (1999) 257–271

Clasification of RAP network using the Lipschitz's semidistance and the measure based on inducement

Sonia Pérez Plaza, Manuel Muñoz Márquez, Fernando Fernández Palacín, and Manuel Berrocoso Domínguez

Universidad de Cádiz
{sonia.perez,manuel.munoz,fernando.fernandez,manuel.berrocoso}@uca.es

Abstract. In this paper the location of the subduction zone between the African and Eurasian plates is studied. A method of classification of functional data is employed in order to determine the limits of this zone. The used data are provided by the RAP network and the method of classification is based on two new similarity measures defined in this work. Furthermore, in order to determine the optimal number of groups in each classification, Silhouette coefficient is employed. Finally, the results obtained are compared to the opinion's experts.

Keywords: FDA, classification, Lipschitz semi-distance, Silhouette coefficient

1 Introduction

The southern part of the Iberian peninsula is over a subduction zone that arise from the convergence of the Euroasian plate and African plate. The limits of this subduction zone are not well determined. In Figure 1 we can see different versions about the location of this zone. In this work we study the problem of location of this zone by using a method of classification.

This region is characterized by a complex seismotectonic pattern and moderate seismic activity associated with the convergence between Africa and Eurasia.

The Andalusian Positioning Network (RAP) is a permanent station network which cover Andalusian area. The stations make a geodetic frame to surveying and cartographic applications.

The seismic movements can be considered as external impulses that generate a displacement in the stations of RAP network. This displacement depends on the situation over the plates. Hence, when we classify the stations, it is important to use a measure that perceive the displacement of the coordinates.

The data employed in this work give the displacement of north and east coordinates between 2011 and 2013. In this paper, these data are considered as functional data.

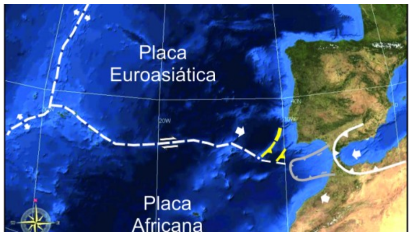


Fig. 1. The limits beetween the Eurasian and the African plate.

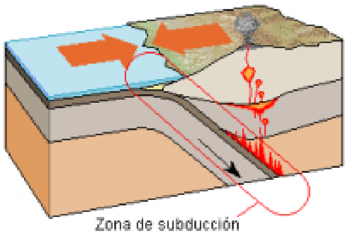


Fig. 2. Subduction zone.

2 FDA

Functional data analysis (FDA) extends the classical multivariate methods when data are functions or curves. According to [1], a functional random variable X is a random variable with values in an infinite dimensional space.

The main source of difficulty when dealing with functional data consists in the fact that the observations are supposed to belong to an infinite dimensional space, whereas in practice one only has sampled curves observed into a finite set of time-points. Indeed, it is usual that we only have discrete observations X_{ij} of each sample path $X_i(t)$ at a finite set of knots.

Because of this, the first step in FDA is often the reconstruction of the functional form of data from discrete observations. The most common solution to this problem is to consider that sample paths belong to a finite dimensional space spanned by some basis of functions ([2]).

An important choice to do when working with functional data is the basis of functions considered.

A basis in FDA is a set of independient functions such that any function can be approximated as a linear combination (of a sufficiently large number) of these functions. Hence, by using a basis, it is possible to aproximate the functional data (of infinite dimension) in a subspace of finite dimension.

The choice of a basis is essential and it must be made by considering how the studied functions are. That is, *Fourier basis* are used with periodic functions, *B-splines basis*, for smooth functions and *Wavelets basis*, for curves that are characterized by numerous local features like peaks or piecewise constants.

3 Similarity measures

Given two functions f and g in $L^2(\tau)$, where $\tau = [T_1, T_2] \subset \mathbb{R}$, the usual distance between these functions is given by:

$$d_{L^2}(f,g) = ||f-g||_{L^2} = \sqrt{(\int_{T_1}^{T_2} (f(t)-g(t))^2 dt)}$$

But this measure seems not to represent the intuitive idea of similitude between curves.

It seems straightforward to see that two parallel curves given by $f(t) = t$ and $g(t) = t + c$ represent two individuals with the same behaviour. Therefore, the distance between f and g should be 0. However, by using the previous distance,

$$d_{L^2}(f,g) = ||f-g||_{L^2} = \sqrt{|T_2 - T_1| * c^2} =$$

$$= c * \sqrt{(T_2 - T_1)} \neq 0$$

In order to avoid this problem, it is possible to use this distance but with the derivatives of the functions. However, in this case, the similitude between curves is not well measured. For this reason, it is necessary to introduced another measure when functional data are employed.

3.1 Lipschitz semi-distance

If we have a function defined between two euclidean spaces (X, d_x) and (Y, d_y), **Lipschitz measure** is defined in [3] as:

$$Lip(f) = \max\{\frac{||f(x)-f(y)||}{||x-y||} : x, y \in X\}$$

Based on this measure **Lipschitz norm** of a function f is defined as follows:

$$||f||_{Lip} = Lip(f)$$

Finally, Lipschitz semi-distance is defined as the semi-distance induced by this norm, i.e.,

$$d_L(f,g) = ||f-g||_{Lip} = sup\{\frac{||(f-g)(x)-(f-g)(y))||}{||x-y||} : x, y \in X\}$$

3.2 Measure based on inducement

Now, we define a similarity measure based in the previously defined Lipschitz semi-distance. In Lipschitz semi-distance, the similarity between curves is given by the maximum in the difference of increasings. In this case, we consider all the relative maximums in the difference of increasings.

Hence, given two functions $f,\, g : \tau \to \mathbb{R}$, the measure based on inducement is defined as follows:

$$d_{est}(f, g) = \sum_{i=1}^{m} \{|h'(x_i)|, x_i \in \tau\}$$

where $h = f - g$ and $x_1, x_2, \ldots, x_m$ are the relative maximums of $|h'|$ in τ.

4 Silhouette coefficient

Silhouette coefficient was proposed by Rousseeuw in 1987 ([4]) in order to provide an evaluation of clustering validity and to select the appropiate number of cluster after a partitioning technique.

Let us first take any object i which is assigned to the cluster A and we compute

$a(i)$ =average dissimilarity of i to all other objects of A.

$$a(i) = \frac{1}{n_A} \sum_{r=1}^{n_A} d(i, a_r), a_r \in A \tag{1}$$

Let us now consider any cluster C which is different from A, and compute $d(i, C)$ =average dissimilarity of i to all objects of C.

$$d(i, C) = \frac{1}{n_C} \sum_{r=1}^{n_C} d(i, c_r), c_r \in C \tag{2}$$

After computing $d(i, C)$ for all clusters $C \neq A$, we select the smallest of those numbers and denote it by $b(i) = min_{C \neq A} d(i, C)$

Now we take the **Silhouette coefficient** at i as

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{3}$$

And finally Rousseeuw define the Silhouette coefficient of the partitioning as

$$s = \frac{1}{n} \sum_{i=1}^{n} s(i) \tag{4}$$

where n is the number of objects in the set.

5 Implementation and results

The data used in this study are the displacement in the north and east coordinates in the GPS stations between the years 2011 and 2013. These data have been modified in two phases: first, we complete the series by using a Kalman filter and, second, we eliminate the values that, according to the experts, are considered as outliers. These modified data have been fitted to Fourier basis.

After this process, we apply a hierarchical grouping procedure to the functional data. By considering the opinion of the experts, the displacements in the plane North-East have been used instead of considering separately the measure of both coordinates. The optimal number of groups has been fixed by considering Silhouette coefficient introduced in this work.

Now, we apply this procedure with Lipschitz semi-distance, with the similarity measure based on inducement, and the distance L^2 over the derivatives.

From this study, we obtain that, with Lispchitz semi-distance, the best classification is obtained, according Silhouette coefficient, for three groups. In this case, we see that a group is only consisting of Melilla (see Figure 4).

In the case where the similarity measure based on inducements is used, the best classification is obtained for two groups (see Figure 3).

Finally, when the usual distance over the derivatives is employed, according to Silhoutte cofficient, the best classification is also obtained for three groups. Therefore, this classification coincides with the classification obtained for Lipschitz semi-distance.

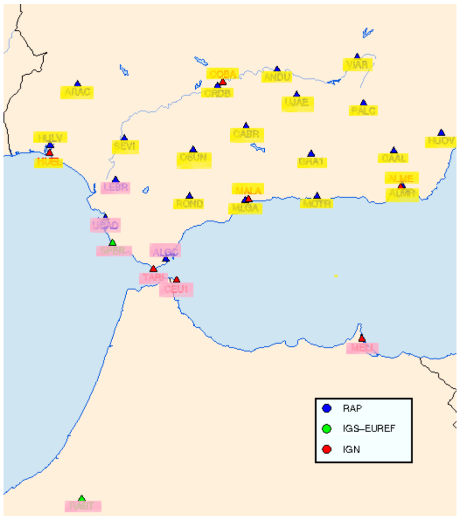


Fig. 3. Classification with d_{est}.

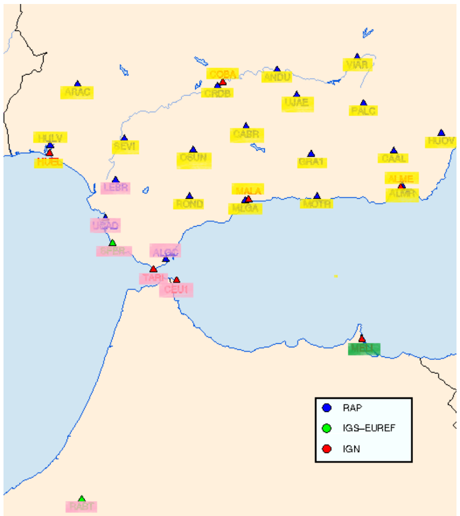


Fig. 4. Classification with d_L.

6 Conclusion

In this work we have studied the location of the subduction zone between the African and Eurasian plates. By using the data of RAP network, we have classified the stations in order to determine the limits of this zone.

In this classification, we have used functional data and for this type of data, we have introduced two new similarity measures. After the process of hierarchical grouping, we conclude that the best classification is the one obtained with the similarity measure based on inducements where two groups are distinguished.

According to the location of the subduction zone estimated by the experts, both groups are placed in both sides of this location and, thus, the obtained results are coherent.

References

1. Ferraty, F., Vieu, P.: Nonparametric functional data analysis. Springer Series in Statistics. Springer, New York (2006)
2. Ramsay, J. O., Silverman, B. W: Functional data analysis. Springer Series in Statistics. Springer, New York, second edition (2005)
3. Weaver, N.: Lipschitz Algebras. World Scientific Publishing Co. Pte. Ltd. (1999)
4. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 20, 53–65 (1987)

Classical Symmetries and conservation laws of the semilinear damped beam equation

J.C. Camacho, M.S. Bruzón, M.L. Gandarias

Departamento de Matemáticas Universidad de Cádiz, Spain
Email: josecarlos.camacho@uca.es,m.bruzon@uca.es,gandarias@uca.es

Abstract. In this paper we study a nonlinear semilinear damped beam equation. We apply the Lie's group theory to make an analysis of the symmetry reductions of the equation. By the multipliers method we look for nontrivial conservation laws via integral formulae.

Key words: Classical Symmetries, Partial Differential Equation, Conservation laws

1 Introduction

In this paper we consider the semilinear damped beam equation:

$$\varDelta \equiv u_{tt} + u_t + u_{xxxx} - \alpha\, u_{xx} = (F(u_x))_x, \tag{1}$$

where $t > 0$, $x \in \mathbb{R}$, $u = u(x,t) : (0,T) \times \mathbb{R} \to \mathbb{R}$ is an unknown function, F is an nonlinear arbitrary function, differentiable, that depends of u_x, and α a positive constant. The equation depends on space, x, and time, t, $u = u(x,t)$ is the deflection of the roadbed and the nonlinear function F models the force beam support.

Equation (1) can be written as a system of differential equations by using a new auxiliary variable v:

$$v = u_x, \qquad u_{tt} + u_t + v_{xxx} - \alpha\, v_x = f(v)\, v_x, \tag{2}$$

where $f(v) = F'(v)$.

In the past decade the theory of group of transformations has been used to arrive to new solutions of partial differential equations (PDEs). We use the classical method for finding symmetry reductions of PDEs, also called Lie group method of infinitesimal transformations, in short when PDEs or ordinary differential equations (ODEs) are invariant under a Lie group of transformations, a reduction transformation exists. In order to obtain all the solutions which are inequivalent with respect to the group it is sufficient to derive the solutions from the optimal system of subalgebras.

Conservation laws appear in many of physical, chemical and mechanical processes, such laws enable us solve problems in which certain physical properties do not change in the course of time within an isolated physical system. The importance of conservation laws also embraces mathematics, for instance, the integrability of a PDE is strongly related with the existence of conservation laws. Furthermore, they can be used to obtain exact solutions of a PDE.

In this paper, we study the classical Lie symmetries of system (2) and we reduce it to systems of ODEs ([15],[1],[6],[7]). The structure is the following: First we find the point transformation group which leaves the system invariant. Next, from the optimal system of subalgebras we find the similarity variables and similarity solutions that reduce the equations to ODEs. In Section 4 we derive by using the multipliers method some nontrivial conservation laws.

2 Symmetry Reductions

Broadly speaking, to apply the classical method for finding symmetry reductions we seek fields of the form

$$v = p(x,t,u,v)\frac{\partial}{\partial x} + q(x,t,u,v)\frac{\partial}{\partial t} + r(x,t,u,v)\frac{\partial}{\partial u} + s(x,t,u,v)\frac{\partial}{\partial v} \qquad (3)$$

that leave the set of solutions of the system (2) invariant.
According to the Invariance Criterion ([15]) we obtain a relationship between the extended variables (u,v,u_x,u_t,v_x,...,u_{xxxx}). Taking into account that these variables are essentially independent, the coefficients in the equation must be equal to zero. This leads us to a system of differential equations in the infinitesimals p,q,r and s so called determining equations.

Symmetries of the system (2)

For system (2) we have that from the determining equations, we only find, finite dimensional algebras.
We obtain the following generators when $F(u)$ is a nonlinear function:

$$w_1 = \frac{\partial}{\partial x},\ w_2 = \frac{\partial}{\partial t},\ w_3 = \frac{\partial}{\partial u},\ w_4 = e^{-t}\frac{\partial}{\partial u}.$$

The optimal system is

$$\{\lambda w_1 + \mu w_2, \lambda w_1 + w_2 + \mu w_3, \lambda w_1 + w_2 + \mu w_4, \lambda, \mu \in \mathbb{R}\}.$$

1. For $\lambda w_1 + \mu w_2$, the symmetry transformation is given by

$$\begin{array}{l} z = \mu x - \lambda t, \\ u(x,t) = h(z) \\ v(x,t) = g(z). \end{array} \qquad (4)$$

Substituting (4) into equation (1) we obtain the ODE

$$\mu^4 \frac{d^4 h}{dz^4} + \left(\lambda^2 - \alpha\,\mu^2 - F'\left(\frac{d\,h}{dz}\right)\right)\frac{d^2 h}{dz^2} - \lambda\frac{d\,h}{dz} = 0. \qquad (5)$$

2. For $\lambda w_1 + w_2 + \mu w_3$, the symmetry transformation given by

$$\begin{aligned} &z = x - \lambda t, \\ &u(x,t) = \mu t + h(z) \\ &v(x,t) = g(z). \end{aligned} \tag{6}$$

leads to the ODE

$$\frac{d^4 h}{dz^4} + \left(\lambda^2 - \alpha - F'\left(\frac{d\,h}{dz}\right)\right)\frac{d^2 h}{dz^2} - \lambda\,\frac{d\,h}{dz} + \mu = 0. \tag{7}$$

3. For $\lambda w_1 + w_2 + \mu w_4$, the symmetry transformation given by

$$\begin{aligned} &z = x - \lambda t, \\ &u(x,t) = h(z) - \mu e^{-t} \\ &v(x,t) = g(z). \end{aligned} \tag{8}$$

leads to the ODE

$$\frac{d^4 h}{dz^4} + \left(\lambda^2 - \alpha - F'\left(\frac{d\,h}{dz}\right)\right)\frac{d^2 h}{dz^2} - \lambda\,\frac{d\,h}{dz} = 0. \tag{9}$$

3 Solutions

Some exact solutions can be obtained from equation (5) when $\mu = 1$: as the derivative of trigonometric, hyperbolic and exponential functions can be expressed in terms of themselves, we can choose F as an algebraic function.

For example $h(z) = \tanh(z)$ is solutions of (9) for

$$F(h) = \frac{1}{2}\left(\lambda + 2\right)\log h - \frac{h\,\left(4\,\alpha - 4\,\lambda^2\right) + 6\,h^2}{4} + C.$$

So, in this case we obtain the solution of (1), $u(x,t) = \tanh(x - \lambda\,t)$

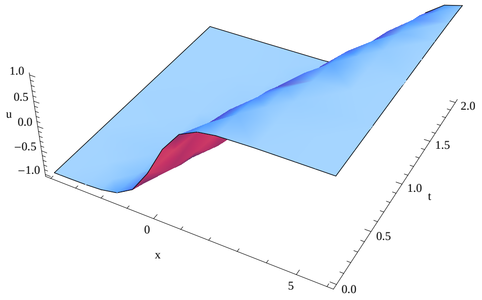


Fig. 1. Solutions $\mathbf{u} = \tanh(x - 2\,t)$

If we take $h(z) = \mathrm{sech}^2(z)$ is solutions of (5) when $\mu = 1$ and for

$$F'(h) = 1 - \frac{\sqrt{1-h}\,\sqrt{h+1}\,\left(12\,h^2 - 2\right)}{3\,h^2 - 2}$$

So, in this case we obtain the solution of (1), $u(x,t) = \mathrm{sech}^2(x - \lambda\,t)$

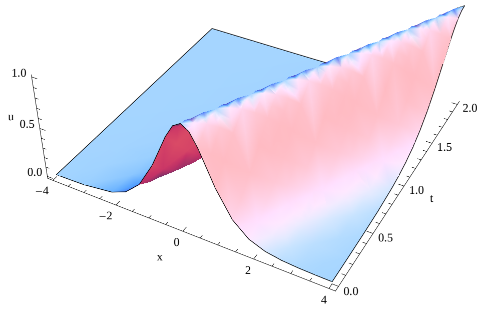


Fig. 2. Solutions $\mathbf{u} = \mathrm{sech}^2(x - 2\,t)$

Some of these solutions are soliton or kink solutions.

4 Multipliers method

Given a PDE a conservation law is a relation of the form

$$D_t(\Phi^t) + D_x(\Phi^x) = 0 \tag{10}$$

where $\mathbf{\Phi} = (\Phi^t, \Phi^x)$ represents the conserved density and flux, respectively, and D_x, D_t denote the total derivative operators with respect to x and t respectively.

In [5] Anco and Bluman gave a general treatment of a direct conservation law method for partial differential equations expressed in a standard Cauchy-Kovaleskaya form in particular for evolution equations

$$u_t = G(x, u, u_x, u_{xx}, \ldots, u_{nx}).$$

The nontrivial conservation laws are characterized by a multiplier λ with no dependence on u_t satisfying

$$\hat{E}[u]\,(\Lambda u_t - \Lambda G(x, u, u_x, u_{xx}, \ldots, u_{nx})) = 0.$$

Here

$$\hat{E}[u] := \frac{\partial}{\partial u} - D_t\frac{\partial}{\partial u_t} - D_x\frac{\partial}{\partial u_x} + D_x^2\frac{\partial}{\partial u_{xx}} + \ldots.$$

The conserved current must satisfy

$$\Lambda = \hat{E}[u]\Phi^t$$

and the flux Φ^x is given by [11]

$$\Phi^x = -D_x^{-1}(\Lambda G) - \frac{\partial \Phi^t}{\partial u_x} G + G D_x \left(\frac{\partial \Phi^t}{\partial u_{xx}} \right) + \ldots.$$

For equation (1) we get the following multipliers.

$$1, e^x, e^x u_x$$

Each multiplier determines a corresponding conserved density and flux:

$$\begin{aligned}
\Lambda = &\ 1, \\
\phi^t = &\ u_t + u, \\
\phi^x = &\ u_{xxx} - \alpha u_x - F(u_x)
\end{aligned}$$

$$\begin{aligned}
\Lambda = &\ e^x, \\
\phi^t = &\ e^x(u_t + u), \\
\phi^x = &\ e^x(u_{xxx} - \alpha u_x - F(u_x)) - \textstyle\int e^x(u_{xxx} - \alpha u_x - F(u_x))dx
\end{aligned}$$

$$\begin{aligned}
\Lambda = &\ \mathrm{e}^x u_x, \\
\phi^t = &\ \mathrm{e}^x u_t u_x, \\
\phi^x = &\ \textstyle\int -\mathrm{e}^x u_x \left(a + \frac{d}{du_x} f\left(u_x\right) \right) du_x + 1/2 \left(2\, u_x u_{xxx} - {u_{xx}}^2 - {u_t}^2 \right) \mathrm{e}^x.
\end{aligned}$$

5 Conclusions

We have applied the Lie classical method to semilinear damped beam equation. Using the characteristic equation, the similarity variables are found. Then, the reduced form of the original nonlinear partial differential equation is obtained as a nonlinear ordinary differential equation. In order to obtain exact solutions we apply a direct method. By this method we have derived some travelling wave solutions. By the multipliers method we have obtained nontrivial conservation laws via integral formulae.

Acknowledgments

The support of DGICYT project MTM2009-11875, FEDER funds and Junta de Andalucía FQM-201 group, is gratefully acknowledged.

References

1. M. Abramowitz, I. A. Stegun, *Handbook of Mathematical Functions*, New York: Dover, 1972.
2. S. Anco, *Conservation laws of scaling-invariant field equations*, J. Phys. A, **36**, (2003), 8623–8638.
3. S. Anco and G. Bluman, *Direct construction of conservation laws from fields equations*, Phys. Rev. Letters, **78(15)**,(1997), 2869–2873.
4. S. Anco and G. Bluman, *Direct construction method for conservation laws of partial differential equations. Part I: Examples of conservation law classifications*, Euro. J. Appl. Math., **15(5)**, (2001), 547–568.
5. S. C. Anco, G.Bluman, *Direct constrution method for conservation laws for partial differential equations Part II: General treatment*, Euro. Jnl of Appl. Math. 2002, 41, 567-585.
6. G.W. Bluman, J.D. Cole, *The general similarity solution of the heat equation*, J.Math. Mech. 1969, 18, 1025-1042.
7. G.W. Bluman, S. Kumei, *Symmetries and differential equations*, Springer-Verlag, 1989.
8. J.C. Camacho, M.S. Bruzón, J. Ramirez, *Classical and Nonclassical Symmetries of the Beam Equation* pp. 2006.
9. B. Champagne, W. Hereman, P. Winternitz, *The computer calculation of Lie point symmetries of large systems of differential equations. Com. Phys. Comm.*, **66**, (1991), 319-340.
10. A.R. Champneys, P.J. McKenna,*On solitary waves of a piecewise linear suspended beam model*, Nonlinearity 10, (1997), 1763-1782.
11. N. Euler, M. Euler, *On nonlocal symmetries, nonlocal conservation laws and nonlocal transformations of evolution equations: two linearisable hierachies.* J. of Nonl. Math. Phys. 2009, 6, 489-504.
12. M.L. Gandarias, M.S. Bruzón, *Classical, and Nonclassical Symmetries of a Generalized Boussinesq Equation*, Journal of Nonlinear Mathematical Physics, Vol. 5, No. 1, 1998, pp. 8-12.
13. A.C. Lazer, P. J .McKenna, *Large Scale Oscillation Behavior in Loaded Asymnmetric Systems.* Ann. Inst. . H-Poincare, Analyse Nonlineaire: Vol. 4, pp.244-274, 1987.
14. P.J. McKenna, W. Walter, *Nonlinear Oscillation in a Suspension Bridge.* Arch. Rational Mech. Anal. 98. pp.167-177, 1987.
15. P.J. Olver,*Applications of Lie groups to differential equations*, Springer-Verlag, 1986.
16. H. Sano,*Finite-dimensional H8 control of flexible beam equation systems*, IMA Journal of Mathematical Control and Information, 19, 477-491. 2002
17. H. Takeda, *Global existence of solutions for higher order nonlinear damped wave equations*, Dynamical Systems and Differential Equations, Proceedings of the 8th AIMS International Conference, Dresden, Germany, 1358-1367 (2011).
18. H. Takeda, S. Yoshikawa, *On the initial value problem of the semilinear beam equation with weak damping I: smoothing effect*, J. Math. Anal. Appl. 401, 244-258 (2013)
19. Takeda, H, Yoshikawa, S: *On the initial data of the semilinear beam equation with weak damping II: asymptotic profiles.* J. Differ. Equ. 253 3061-3080 (2012).
20. H. Takeda, *Large time behavior of global solutions of the semilinear damped beam equation with slowly decaying data* , J. Math. Anal. Appl. 423, No. 2, 898-912 (2015). ISSN 0022-247X

Automatic Detection of Metamorphic Relations: A Challenge for WS-BPEL

C. Castro-Cabrera and I. Medina-Bulo

UCASE Software Engineering Research Group,
University of Cadiz, Spain {maricarmen.decastro,inmaculada.medina,}@uca.es
https://ucase.uca.es/

Abstract. Web services have nowadays great impact on society due to numerous internet transactions existing. WS-BPEL is a Bussiness Process Enterprise Language that allows implement compositions as web services. This type of software requires to be tested to avoid errors and fatal consequences. In a previous work, the authors proposed to apply the Metamorphic Testing technique to WS-BPEL compositions through a particular architecture. That approach has some steps as the identification and implementation of properties to be used. This paper focuses on the composition analysis for obtaining information to automate that step.

Keywords: Metamorphic Testing, WS-BPEL, oracle, Metamorphic Relations, follow-up test cases

1 Introduction

Web Services Business Process language, WS-BPEL 2.0 [9] was standardized at the request of some TIC companies (HP, IBM, Oracle, Microsoft, etc.). This language allows us to develop a new Web Service (WS) designing more complex business processes from pre-existing WS, and there is a widely support software for them. However its development has not gone along with improvements on testing techniques to this type of software [1]. A deficient testing in a system could cause errors with negative consequences both economical and also human. Consequently, good testing methods to test correctness of compositions are required. Progresses in this sense are described in [10].

Metamorphic Testing (MT) [5] is a software testing technique using *metamorphic relations* (MRs). MRs are existing or expected relations defined on a set of inputs and their corresponding outputs for multiple executions of a function under test. The underlying concept is simple and its automation is not difficult. In fact, it has proved successful in testing and improving the quality of traditional imperative programs [14].

Regarding the cost effectiveness of MT, Zhang [13] conducted an experiment where the fault detection capabilities and time cost of MT were compared to the standard assertion checking method. Results showed that MT has the potential to detect more faults than the assertion checking method.

This paper discusses how to use MT to test WS compositions in WS-BPEL. Although MT has not been previously applied to this area, promising results have been obtained in a number of different applications. A component diagram for a testing framework implementing this approach is included as well as alternatives to automate the analysis step for the MR identification and implementation.

The structure of the rest of the paper is as follows: An introduction about metamorphic testing and WS-BPEL language are respectively shown in Section 2 and Section 3. Section 4 includes the particular architecture. Section 5 describes different alternatives for the analysis step and, finally, Section 6 presents the conclusions and future work.

2 Metamorphic testing

There are properties associated to some functions or applications, such that, if the inputs are changing (i.e.increase or decrease in a quantity), it should be possible calculate the new output through the output generated from the original input, without need to run the program. Therefore, if the new input is executed by the program, the output must be the same that had been calculated previously. The operations to apply to the inputs constitute properties, which relate the initial tests cases with the follow-up tests cases, they are called Metamorphic Relations (MR). MT is based on this notion and easily carried out in practice. The original test cases and their corresponding follow-up test cases are constructed based on these MRs. Both of them are executed using the program under test, to verify it. If any test case does not satisfy a MR, an error is detected.

For instance, given a program implementing the arithmetic mean of a set of numbers, permuting the order of the elements should not affect to the mean calculation; If other operations are applied such as, multiplying or increasing each value by a number, the resulting mean (follow-up test case output) should be easily predicted, multiplying or increasing (respectively) the original mean by that number. If the different outputs for their corresponding inputs are not as expected, then there must be a error in the implementation. Other example, based on lists of numbers it is showed in figure 1. The program *f* plays the role of the *reverse order* function and *t*, multiply by 3, the MR:

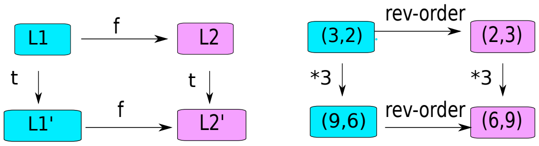


Fig. 1. Metamorphic Relation Example

3 WS-BPEL Composition Language

WS-BPEL is a programming language based on XML that is used to generate business processes from services defined previously. The resulting business process can be then reused as a WS in higher level compositions. A WS-BPEL composition contains four sections including declarations of: the relationships to the external partners, the variables, handlers, and description of the business process behavior. The major building blocks in WS-BPEL are the *activities*. Furthermore, WS-BPEL provides concurrency and synchronization primitives. Here is an example:

```
<flow>  ← Structured activity
 <links>  ← Container
  <link name="checkFl-BookFl"/> ← Element
 </links>
 <invoke name="checkFlight" ... > ← Basic activity
  <sources>  ← Container
   <source linkName="checkFl-BookFl"/> ← Element
  </sources>
 </invoke>
 <invoke name="checkHotel" ... />
 <invoke name="checkRentCar" ... />
 <invoke name="bookFlight" ← Attribute ...>
  <targets>  ← Container
   <target linkName="checkFl-BookFl" />
  </targets>
 </invoke>
</flow>
```

4 MT implementation and architecture

Firstly, it is mentioned a general implementation of MT, presented in [4]. The sequence is: Choose the initial test suite, select adequate MRs, generate the follow-up test suite applying MRs to initial test suite, execute the program with initial and follow-up test cases, compare the result and finally, improve the program correcting the detected errors, select new test cases and/or new MRs enhanced to successive iterations.

The goal is to implement MT to test WS-BPEL compositions. Therefore, it is necessary to take into account the peculiarities of this language and the compositions. Figure 2 describes the particular architecture lightly improved to include MuBPEL [6], a mutation tool to validate the technique. MuBPEL is a mutation testing tool for WS-BPEL 2.0. It can be used to evaluate the quality of a test suite by checking if it can tell apart a mutant from the original program. Mutants are slightly modified (mutated) versions of the original program in which a single syntactical change has been made: for example, ”$4 + 5$” may have been changed to ”$4 - 5$”or”$4 + 6$”. In this way, MuBPEL is used to validate the MT technique. If a test case does not satisfy a MR, an error is detected (a mutant is killed).

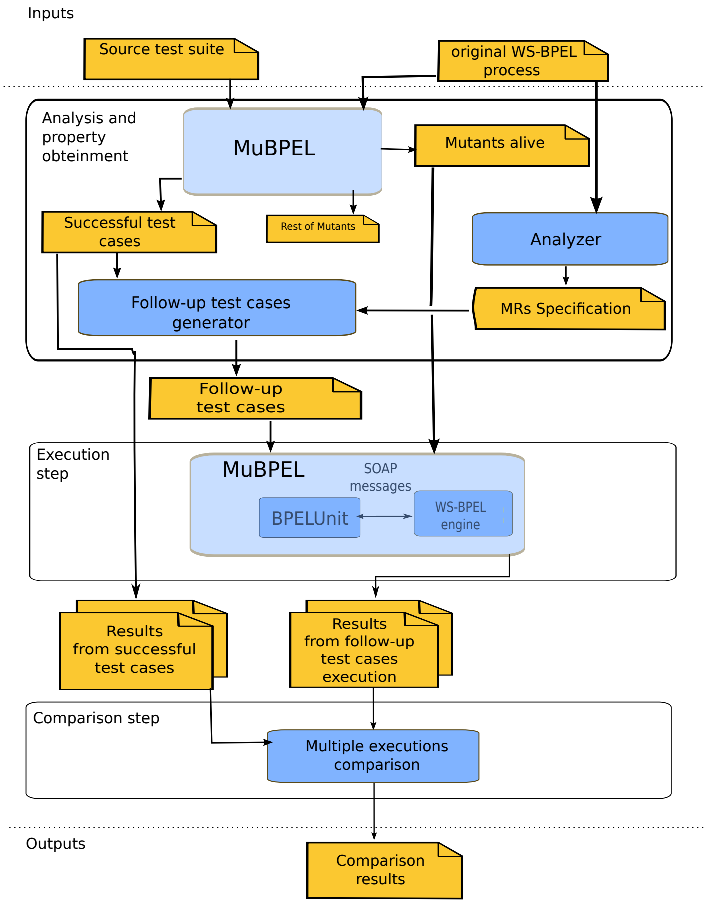


Fig. 2. Architecture to apply MT in WS-BPEL

A prototype was built following this arquitecture and analysing some compositions to obtain the properties. The MRs were designed and implemented by hand for each composition. Besides in previous works [2] and [3] that prototype was applied to some case studies with promising results. However, both the composition analysis and the obtained MRs were processed by hand.

Some MT applications on the literature are about problems or program whose properties are known previously. For instance, there are numerous works about machines learning applications or mathematical functions [8], [12]and [7]. This step is a challenge for WS-BPEL compositions. Several alternatives are proposed in the following section.

5 Alternatives to analyse the composition

Our goal is scanning every composition and to extract relevant information that assesses the framework to design and implement MRs. So somehow, this step makes easy generating MRs. For this purpose, we offer some possibilities (not exclusive):

1. Up the bussiness level (bussines rules)
2. Use Takuan [11] (invariant generator to WS-BPEL compositions)
3. Create a new analyzer to extract other information

The first option does not appear simple or applicable (The usual approach is the inverse step). The natural step goes in the inverse order. This requires working on the bussiness level to extract the composition behaviour, implementing properties and applying them.

With respect to the second option, Takuan is an open-source WS-BPEL dynamic invariant generator which can infer invariants from WS-BPEL process definitions. It generates relevant information, but perhaps too simple for our purpose. For instance, there are some invariants from Loan Approval composition as follow:

$$\text{loanApprovalProcess.\_process1\_sequence1:::EXIT}$$
$$\text{request.amount} == \text{orig(request.amount)}$$
$$\text{request.amount one of } \{ \ 1500,\ 15000,\ 150000 \ \}$$
$$\text{risk.level one of } \{ \ \text{"high", "low"} \ \}$$

It would be necessary to combine them to obtain information useful for MR implementation. So, they could be complemented with the third option, developing a new application to extract information to asses the MR implementation. The goal of the third option is to locate key values and expressions in every composition to determine the properties to build. For example, a numeric constant in a condition could lead to an arithmetic property or a logical expression could lead to a logical relation between some test cases values.

6 Conclusions and future work

WS-BPEL business processes are considerably increasing in last years. For this reason, it is important the development of techniques that allow to test this type of software. Due to the language nature, specific to WS, it is necessary to implement alternatives tecniques to test this kind of compositions.

In addition, MT has been implemented in different languages efficiently and applications have been tested on various study fields such as medicine or bioinformatics. Actually, more than 80 papers have been published about this subject. Selection of adequate MRs is an important issue to this technique, so we ought to consider the problem context and the structure of the program under test.

A testing framework architecture to apply MT in WS-BPEL compositions is being built. Further some possibilities to guide the property identification and implementation have been presented. The future work includes to implement all steps and compare with other techniques.

References

1. Bozkurt, M., Harman, M., Hassoun, Y.: TR-10-01: testing web services: A survey. Tech. Rep. TR-10-01, King's College, London (2010)
2. Castro-Cabrera, C., Medina-Bulo, I.: Application of Metamorphic Testing to a Case Study in Web Services Compositions, Communications in Computer and Information Science, vol. 314, pp. 168–181. Springer Berlin Heidelberg (2012), http://dx.doi.org/10.1007/978-3-642-35755-8_13
3. Castro-Cabrera, C., Medina-Bulo, I., Camacho, A.: Metasearch services composition in ws-bpel - an application of metamorphic testing. In: ICSOFT (2012)
4. Castro-Cabrera, M.d.C., Camacho-Magriñàn, A., Medina-Bulo, I., Palomo-Duarte, M.: Una arquitectura basada en pruebas metamòrficas para composiciones de servicios ws-bpel. In: Actas de las VII JCIS. pp. 9–22. Servizo de publicaciòns da Universidade da Coruña, A Coruña, España (Sep 2011)
5. Chen, T.Y.: Metamorphic testing: A new approach for generating next test cases. HKUSTCS98-01 (1998)
6. García-Domínguez, A., Estero-Botaro, A., Medina-Bulo, I.: MuBPEL: una herramienta de mutación firme para WS-BPEL 2.0. In: Actas de las XVI JISBD (2012)
7. Mahmud, S., Elahe, M.F., Jahan, H.: Verifying mathematical function using metamorphic testing. International Journal of Research in Computer Engineering & Electronics 3(5) (2014)
8. Murphy, C., Kaiser, G., Hu, L., Wu, L.: Properties of machine learning applications for use in metamorphic testing. In: Proc. of the 20th International Conference on SEKE (SEKE). pp. 867–872 (2008)
9. OASIS: Web Services Business Process Execution Language 2.0. http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html (2007), Organization for the Advancement of Structured Information Standards
10. Palomo-Duarte, M.: Service composition verification and validation. In: Jonathan Lee, S.P.M., Liu, A. (eds.) Service Life Cycle Tools and Technologies: Methods, Trends and Advances, pp. 200–219. IGI Global (2011)
11. Palomo-Duarte, M., Garcia-Dominguez, A., Medina-Bulo, I.: Automatic dynamic generation of likely invariants for ws-bpel compositions. Expert Systems with Applications 41(11), 5041 – 5055 (2014), http://www.sciencedirect.com/science/article/pii/S095741741400061X
12. Xie, X., Ho, J.W., Murphy, C., Kaiser, G., Xu, B., Chen, T.Y.: Testing and validating machine learning classifiers by metamorphic testing. Journal of Systems and Software 84(4), 544 – 558 (2011), the Ninth ICQS
13. Zhang, Z.Y., Chan, W.K., Tse, T.H., Hu, P.F.: An experimental study to compare the use of metamorphic testing and assertion checking. Journal of Software 20(10), 2637–2654 (2009)
14. Zhou, Z.Q., Huang, D.H., Tse, T.H., Yang, Z., Huang, H., Chen, T.Y.: Metamorphic testing and its applications. In: Proceedings of the 8th ISFST(ISFST 2004). Software Engineers Association (2004)

Semandal: Extracting knowledge from city councils *

Daniel Albendín Moya[1], Gonzalo A. Aranda-Corral[1], Joaquín Borrego-Díaz[2], and Angel Cantó Vicente[1]

[1] Universidad de Huelva. Department of Information Technology.
Crta. Palos de La Frontera s/n. 21819 Palos de La Frontera. Spain
[2] Universidad de Sevilla. Department of Computer Science and Artificial Intelligence.
Avda. Reina Mercedes s/n. 41012 Sevilla. Spain

Abstract. One of the main goals of Semantic Web is to make all available information machine-readable and understood by other machines. For these, ontologies are key elements that will enable us to exploit all the advantages. Ontologies try to model the world in order to represent all web information. But, for general purposes, this is too broad and ambitious goal for only a single ontology or platform. In order to easy the creation of the ontology we reduce the scope only to news extracted from city councils web pages.
In this paper, we present a project where information and data are collected from webs and processed by means of Formal Concept Analysis to align it to an "ad-hoc" ontology.

Keywords: Knowledge extraction, Formal concept analysis, Semantic Web

1 Introduction

As the W3C establish in its web page: " *The Semantic Web[2] provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries...* In order to achieve these goals, traditional web should be translated into data or machine readable documents, located by URIs, and they are also be related to others. Semantic Web technologies can be used in a variety of application areas. Our interest is focused on information integration, where information from different sources can be organized to enhance its access, organization, etc.

Nowadays, Open Government Data (OGD) is other important emerging trend that merges the Open Data foundations with Public entities. OGD are data produced by or commissioned by government and its intended for freely used, reused and redistributed by anyone. It has great benefits as transparency, social and commercial value and let participatory Governance.

* Partially supported by TIC-6064 Excellence project (Junta de Andalucía) and TIN2013- 41086-P (Spanish Ministry of Economy and Competitiveness), cofinanced with FEDER funds

Semandal is a platform that tries to apply all concepts about Semantic Web and Open Government Data on the municipalities of Andalucía, Spain. This scope was chosen to reduce the dimensions of vocabularies, ontologies, and, even, databases. Semandal extracts the information from the traditional web pages and transform it into knowledge and republish it by means of an API and a structured format (machine readable). Even this, Semandal provides a mobile app to let user to access this information, but this is out of scope of this paper.

To transform information into knowledge we use Formal Concept Analysis[1] (FCA). FCA is a non supervised clustering technique which, from formal definitions, can construct concept lattices and set of rules which represents information.

Semandal's architecture is depicted in fig 1, where main modules can be distinguished and will lead the structure of our paper (only two firsts): Extraction, Classification and re-publishing.

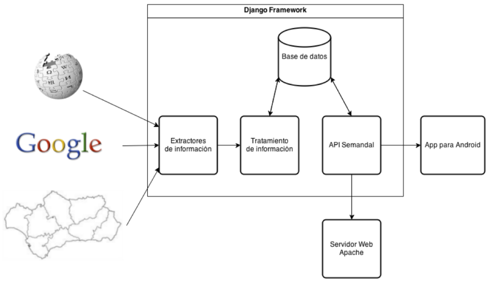


Fig. 1. Semandal's Architecture

1.1 Formal Concept Analysis

A useful bridge between Semantic Web and knowledge extraction could be *Formal Concept Analysis* (FCA) [1].

According R. Wille, FCA mathematizes the philosophical understanding of a concept as a unit of thought composed by two parts: the extension and the intension. The extension covers all objects (documents) belonging to this concept, while the intension comprises all common attributes (tags) valid for all the objects under consideration. It also allows the computation of concept hierarchies out of data tables, and it is also used for ontology mining from folksonomies. Several applications from FCA to tagging, folksonomies and semantic tasks have been developed (see [6]).

We represent a formal context as $M = (O, A, I)$, which consists of two sets, O (objects) and A (attributes) and a relation $I \subseteq O \times A$. Finite contexts can be represented by a 1-0-table (representing I as a Boolean function on $O \times A$). Basic FCA logical expressions are implications between attributes, that is, a pair of sets of attributes, written as $Y_1 \to Y_2$, which is true with respect to

$M = (O, A, I)$ and is defined as follows. A subset $T \subseteq A$ *respects* $Y_1 \to Y_2$ if $Y_1 \nsubseteq T$ or $Y2 \subseteq T$. We say that $Y_1 \to Y_2$ holds in M ($M \models Y_1 \to Y_2$) if for all $o \in O$, the set o' respects $Y_1 \to Y_2$. In that case, we say that $Y_1 \to Y_2$ is an implication of M.

Every implication has also associated some properties, e.g. support. Support is defined as the number of objects that contain all attributes from Y_1 and holds the implication.

2 Knowledge extraction and integration

One of the main goals of Semantic Web is information integration from different sources and, in our case, we start with a file of municipalities, obtained from INE[3], written in XLS format. From these, we develop software to connect to several APIs, from Google, Wikipedia, etc.. From Google Search we locate the web page address for all municipalities and, in some cases, where disambiguation is needed, we use Wikipedia to achieve it. Also we used other APIs to extract the geolocation, population, extension, etc...

After all web addresses were collected we found a big challenge, news extractions. All web pages were organized in really different ways which makes this task really hard. Based on most common patterns, we found up to 7 different clusters of similar organizations, and we built an extractor for each one.

Some of them, based on page's structure or news content, can extract the category associated to news, classified by publishers (supposed to be valid). Other extractors were not able to do it. From these, we obtained a set of classified and other unclassified news, which are quite appropriate to apply machine learning techniques.

3 Categories selection

From the set of classified news, we study the set of categories to remove any possible mistake.

First task was to remove typos, plurals or abbreviations, grouping words based on Levenshtein distance which are closer than 3, obtaining a set of categories not too big and working.

In order to semantize as much content as possible, we searched for a hierarchy or ontology where map our set of categories soundly, but we did not find any one which fits with our needs about municipalities news. To solve this, we created an "ad hoc" categories' hierarchy (fig. 2) where also some of extracted categories are removed (out of meaning in our scope) and merging some of them with the same meaning. Needed super-classes are added too.

From this hierarchy, we reclassify all news applying it: aligning categories to hierachy's concepts and adding all super-classes.

[3] Instituto Nacional de Estadística.(National Institute of Statistics)

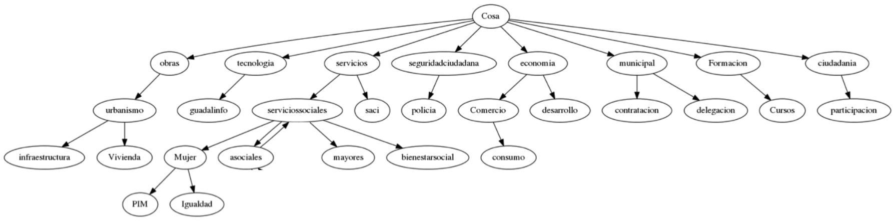


Fig. 2. Hierarchy of categories

4 Classification of information

Now, we are able to classify all news which were previously unclassified from its own content and other classified news. From these news, we planed to build a classifier based on words contained into the text. Firstly, we remove non significant words, as prepositions, articles, own names, numbers, etc...

Secondly, we calculated the relative weights of each word in the categories where they appear to obtain the significance of it into the category. After, we calculated the average of weights for each category and, if this number is high means that the word is too common in all categories and we should remove it. Other considered option, but with similar results we based on [3]

If we create a graph with resulting words connecting them and categories, setting each edge's weight to calculated frequency value. This graph is shown in fig. 3.

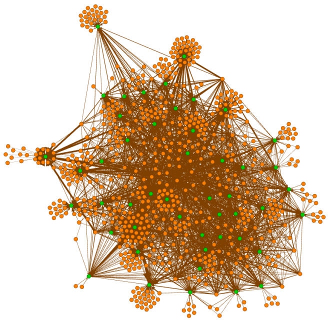


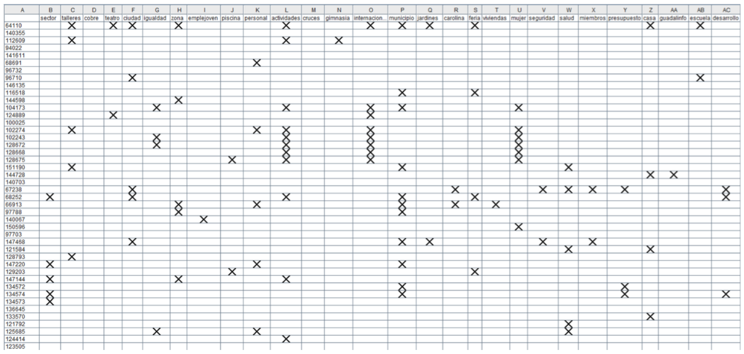
Fig. 3. Categories - Words

As we can see, there are words that define categories fairly well, but others remain poorly relevant. Nevertheless, we think this set of words is good enough to implement the classifier.

4.1 Formal Concept Analysis

Construction of classifier we made by means of a set of rules obtained applying Formal Concept Analysis technique. As above, we need to build a formal context to get a concept's lattice, as first step, which we can take as our first emergent ontolgy prototype about news, and , finally, a set of association rules which will be the responsible to classify news.

The formal context considered classified news as objects and relevant words and extracted categories as attributes. This will infer a set of attribute rules (with support and confidence) which, applied to unclassified news it will infer a new set of categories. Each new set of categories will be the new assigned to each news. An example of this context is shown in fig. 4.



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
	sector	talleres	cobre	teatro	ciudad	igualdad	zona	emplejoven	piscina	personal	actividades	cruces	gimnasia	internacion...	municipio	jardines	carolina	feria	viviendas	mujer	seguridad	salud	miembros	presupuesto	casa	guadalinfo	escuela	desarrollo
64110		X		X	X		X				X			X	X	X		X							X		X	
140355																												
112609		X									X		X															
94022																												
141611																												
68691										X																		
96732																												
96710					X																						X	
146135																												
116518															X			X										
144598							X																					
104173						X					X			X	X					X								
124889				X										X														
100025																												
102274		X								X	X			X						X								
102243						X					X			X						X								
128672						X					X			X						X								
128668											X			X						X								
128675									X		X			X						X								
151190		X													X							X						
144728																									X	X		
140703																												
67238					X												X				X	X	X	X				X
68252	X				X						X				X			X										X
66913							X			X					X		X		X									
97788							X								X													
140067								X																				
150596																				X								
97703																												
147468					X										X	X					X		X					
121584																						X			X			
128793		X																										
147220	X									X					X													
129203									X									X										
147144	X						X				X																	
134572															X									X				
134574	X														X									X				X
134573	X																											
136645																												
133570																									X			
121792																						X						
125685						X				X												X						
124414											X																	
123505																												

Fig. 4. Formal context

5 Experiments

For our experimentation, we had to choose a proper subset of news, since the total number of news is huge. We prepared some experiments to find out an affordable amount of size or time in order to obtain sound results.

We built 3 contexts, A, B and C, with different number of objects and attributes obtaining a number of rules which grow exponentially, as shown in table 1. In order to reduce this number of rules, we only considered as valid rules (Classif) that ones which have a *support* > 0 and have, at least, one category on the right side of the rules.

To build the expert system to generate new categories, we translate the rules into a CLIPS[4] format and we run it by means of a Jess[5] program, also developed by us.

Finally, we test the expert systems with 2 random news to check its soundness. News are in spanish, because of all database is focused on Spanish municipalities.

Context	Attributes	Objects	Rules	Classif.
A	145	133	7412	2650
B	208	258	26344	9813
C	257	372	66910	53898

Table 1. Size of experiments

[Noticia 1] "*El novillero de Écija Antonio David, proclamado triunfador de la V feria de novilladas de promoción la granada de plata*" [4]	[Noticia 2] "*El ayuntamiento da luz verde para la construcción de otras 75 viviendas protegidas*"[5]
A: Turismo, Juventud	A: Vivienda
B: Turismo, Cultura	B: Turismo, Servicios sociales
C: Festejos	C: Servicios sociales, Obras

6 Remarks and future works

Building a platform for extracting and processing information is a really hard task for machines, even when it is limited in scope, that still needs some work. Nonetheless, we have shown, in this paper, that FCA could be a useful tool to build classifiers that would allow us to transform extracted information into knowledge in an affordable way with soundness.

There are still other tasks within this platform which are really relevant for final success, as the acquisition of municipality's general information, not only news, it means, information about organizational structure, contacts, plenary sessions, etc... . From knowledge point of view, next step should be a deep work on words, using networks of concepts, as WordNet, to align them to the network and integrate in some RDF Open Data catalog.

References

1. B. Ganter and R. Wille. Formal Concept Analysis - Mathematical Foundations. Springer, 1999.
2. T. Berners-Lee, J. Hendler, O. Lassila, *The Semantic Web*, Scientific American, 284(5), pp. 34–43, 2001
3. Gerard Salton and Michael J. McGill. 1986. Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York, NY, USA.
4. CLIPS (2009). CLIPS: A tool for building expert systems.
5. Ernest Friedman Hill. 2003. Jess in Action: Java Rule-Based Systems. Manning Publications Co., Greenwich, CT, USA.
6. R. Jaschke, A. Hotho, C. Schmitz, B. Ganter, and G. Stumme. Discovering shared conceptualizations in folksonomies. Journal of Web Semantics, 6(1):pp 38-53, 2008.

Dichotomous sets of implications and directness

E. Rodríguez-Lorenzo, P. Cordero, M. Enciso, and A. Mora

University of Málaga, Andalucía Tech, Spain,
e-mail: {estrellarodlor,amora}@ctima.uma.es, {pcordero,enciso}@uma.es

Abstract. In this paper, we study the directness property of implications in formal concept analysis. We show how dichotomously split the implications in two subsets according to their premise closure. Thus, we define a new directness paradigm strongly based on a separated treatment of the two implication subsets and how to compute the proposed dichotomous direct implicational system from a set of implications.

1 Introduction

In Formal Concept Analysis (FCA), closure operators are one of the most basic notions. These operators allow to solve important exponential problems in different areas such as formal concept analysis, AI, databases, etc.

Moreover, closure operators are directly related to implications. K. Bertet points to a good direction in [3] where Implicational Systems (IS) are highlighted as convenient tools to handle a closure system. So, it makes sense that the search for the efficiency in the set closure computation is a major challenge [1].

In [4] K. Bertet et al. establish specific properties to achieve the mentioned goal. The properties that they considered is the directness and optimality, that is, the computation of the closure of an attribute set can be performed in one traversal of the implicational set and none implication can be removed without losing this property. Then, progressing this line, K. Adaricheva et al. [1] propose the so called *D-basis* as a subset of the basis proposed in [4], which is direct as well and has less number of implications.

In this paper, we are working in the design of new IS definitions suitable to describe closure system. We pay attention to the closure of each premise to make a separate treatment of those ones whose closure is the total set of attributes to provide an improvement in the performance of closure methods. Moreover, the premise of these implications fit exactly with the notion of keys in database and generators in FCA.

Here, we propose a new definition of IS named dichotomous IS whose main characteristics is the separate treatment of implications depending on the closure of its premise. Moreover, we introduce the notion of direct dichotomous basis (DD-basis) and illustrate its advantages.

2 Formal Concept Analysis

Formal Concept Analysis (FCA) is a formal framework oriented to data analysis and knowledge discovering. FCA extracts knowledge from the information presented in a formal context providing equivalent ways of representation of knowledge: concept lattices and implicational systems. We focus on the second ones because they can be managed and depurated by using logic.

The relationship between a set of objects and a set of attributes are described using a formal context as follows:

Definition 1. *A formal context is a triple* $\mathbf{K} = (G, M, I)$ *where* G *is a finite set whose elements are named objects,* M *is a finite set whose elements are named attributes and* $I \subseteq G \times M$ *is a binary relation. Thus,* $(g, m) \in I$ *means the object* g *has the attribute* m.

The concept of implication is a central point in this work together with a special case of implication introduced as follows:

Definition 2. *Let* $\mathbf{K} = (G, M, I)$ *be a formal context and* $A, B \in 2^M$. *The implication* $A \to B$ *holds in* $\mathbf{K}$ *if every object* $g \in G$ *satisfies the following:* $(g, a) \in I$ *for all* $a \in A$ *implies* $(g, b) \in I$ *for all* $b \in B$.

Given a subset $X \subseteq M$*,* X *is a key if the implication* $X \to M$ *holds in* $\mathbf{K}$*. An implication* $X \to Y$ *such that* X *is a key will be named a key implication.*

For this work, we only present a brief summary (only the rules) of the Simplification Logic ($\mathbf{SL}_{\text{FD}}$), an equivalent logic to Armstrong's Axioms [2] but more adequate to develop automated method for implications. See [5, 8] for a more detailed presentation of $\mathbf{SL}_{\text{FD}}$ its semantic, and how remove redundancy and compute closures using directly the $\mathbf{SL}_{\text{FD}}$.

It is assumed to be familiar with notions of the derivation of an implication from an IS, the semantic entailment and the equivalence between two ISs.

Definition 3 (Rules of $\mathbf{SL}_{\text{FD}}$).
Reflexivity as axiom scheme and the following inference rules named fragmentation, composition and simplification are considered in $\mathbf{SL}_{\text{FD}}$.

$$\texttt{[Ref]}\ \ \frac{}{A \to A} \qquad \texttt{[Frag]}\ \ \frac{A \to BC}{A \to B} \qquad \texttt{[Comp]}\ \ \frac{A \to B,\ C \to D}{AC \to BD} \qquad \texttt{[Simp]}\ \ \frac{A \to B,\ C \to D}{A(C\text{-}B) \to D}$$

Definition 4. *Let* $\Sigma \subseteq \mathcal{L}_S$ *be an IS and* $X \subseteq S$. *The closure of* X *wrt* Σ *is the largest subset of* S*, denoted* X^+_Σ*, such that* $\Sigma \vdash X \to X^+_\Sigma$.

Finally, the notion of key, which plays a central role in this paper, can be characterized in terms of ISs. Thus, we are going to say that an attribute set X is key with respect to an IS Σ if it is a key with respect to any model of Σ.

Proposition 1. *Let* $\Sigma \subseteq \mathcal{L}_S$ *be an IS and* $X \subseteq S$. *The following conditions are equivalent:*

1. X *is a key with respect to any model of* Σ.
2. $\Sigma \vdash X \to S$.
3. $X^+_\Sigma = S$.

3 Dichotomous set of implications

Fist, we emphasize that although some closure algorithms to solve these problems have a linear cost, due to its exhaustive use in some NP algorithms, a minor gain in the closure performance entails a major advantage for these complex methods.

Now, the goal is to remain the advantage of the directness which establishes that the closure of an attribute set may be computed with just one traverse of the set of implications. Thus, our approach is to provide an alternative direct IS such that it can be obtained with less cost.

The study of directness demands the design of closure operators for attribute sets. Thus, other authors who have studied different kinds of direct basis define closure operators for these bases: the direct-optimal basis [4] and the D-basis [1].

Now, we introduce the notion of dichotomous set of implications and a twofold operator suitable for its management.

Definition 5 (Dichotomous set of implications). *A pair of implicational sets $\langle \Sigma^*, \Sigma^k \rangle$ is named a dichotomous set of implications if all $A \to B \in \Sigma^k$ are key implications.*

We define the σ operator for dichotomous ISs as a composition of two underlaying operators: $\sigma_{\langle \Sigma^*, \Sigma^k \rangle} = \kappa_{\Sigma^k} \circ \pi_{\Sigma^*}$[1] where

$$\kappa_{\Sigma}(X) = \begin{cases} M & \text{if } A \subseteq X \text{ for some } A \to B \in \Sigma \\ X & \text{Otherwise} \end{cases}$$

Assuming $\langle \Sigma^*, \Sigma^k \rangle$ being a dichotomous set of implications, we have that $\sigma_{\langle \Sigma^*, \Sigma^k \rangle}$ is isotone and extensive. The directness property can also be considered in this framework by means of the idempotence of the operator $\sigma_{\langle \Sigma^*, \Sigma^k \rangle}$.

Definition 6. *A dichotomous set of implications $\langle \Sigma^*, \Sigma^k \rangle$ is said to be direct if $\sigma_{\langle \Sigma^*, \Sigma^k \rangle}$ is a closure operator.*

The following proposition leads the way to characterize a direct dichotomous set of implications and the next corollary characterizes the directness in our dichotomous approach.

Proposition 2. *Let $\langle \Sigma^*, \Sigma^k \rangle$ be dichotomous set of implications. For all $n > 0$ we have that $\sigma^n_{\langle \Sigma^*, \Sigma^k \rangle} = \kappa_{\Sigma^k} \circ \pi^n_{\Sigma^*}$.*

Corollary 1. *A dichotomous set of implications $\langle \Sigma^*, \Sigma^k \rangle$ is direct if and only if Σ^* is a direct IS.*

In this section we have focused on the directness property. Nevertheless, this is not an isolate property and it is usually related to other properties leading to the notion of basis, which constitutes the main issue of the following section.

[1] π_Σ is the operator defined in [3]: $\pi_\Sigma(X) = X \cup \{b \in B | A \subseteq X \text{ and } A \to B \in \Sigma\}$

4 DD-basis

In this section we are interested in the improvement of ISs demanding some optimality or minimality properties. The property related to the notion of basis is minimality, which means that if any implication is removed from the set of implications, it is not equivalent to the initial one. But in this paper, the property that we need is the optimality: an IS is optimal if there is not an quivalent IS with less size, where the size is: $\|\Sigma\| = \sum_{A \to B \in \Sigma}(|A| + |B|)$.

The above properties are used in the literature to introduce different notions of basis, but another property must be introduced. The well-known Duquenne-Guiguess basis was introduced in [7] being a minimum (there is not another equivalent IS with less cardinality), but no direct, IS. In this paper we are interested in those basis strongly related with the directness property. In this way, Bertet et al. propose the following definition adding the optimality property.

Definition 7 (Direct-optimal basis [3]). *An IS Σ is named a direct-optimal basis if Σ is direct and any other equivalent direct IS has a greater size.*

Alternatively, K. Adaricheva et al. [1] introduce another basis related to the directness property and taking into account the order of the implications: the D-basis (see [1] for more details). A relevant issue in the area of direct basis is the cost of its computation. That is why we provide here an alternative direct basis definition in the framework of the dichotomous ISs.

Definition 8 (DD-basis). *Let $\langle \Sigma^*, \Sigma^k \rangle$ be a dichotomous set of implications, we say that it is a dichotomous direct basis, briefly DD-basis, if it is minimal and $\sigma_{\langle \Sigma^*, \Sigma^k \rangle}$ is a closure operator.*

The main advantage of the proposed DD-basis with respect to both alternative direct bases, is that our approach reduces the size of the subset of implications withstanding the exponential cost of the basis construction process, as the following section shows. This reduction comes from the removal of the key implications in the exponential task.

5 A method to compute the DD-basis

Most of the knowledge discovering methods in FCA returns a Duquenne-Guigues basis. In this section we focus on the design of an efficient method to compute a DD-basis equivalent to a given Duquenne-Guigues basis.

The use of the dichotomous set of implications is motivated by the idea of reducing the input of the costly task in the basis computation method. Corollary 1 establishes that a dichotomous set of implications is direct if and only if the first component is direct. Thus, we begin this section with the description of the transformation method of the first component Σ^* into the corresponding equivalent direct-optimal basis by executing a modification of the algorithm we presented in [9]. Now, we briefly describe this algorithm. In a first stage the method simplifies implications with redundant attributes in both left and right hand sides, transforming Σ^* into its equivalent reduced one, denoted Σ_r^*, requires just the application of the rules of the $\mathbf{SL}_{\text{FD}}$.

Later, the algorithm exhaustively applies the inference rule called *strong simplification*, **[sSimp]**, that covers directness without losing reduceness:

$$\texttt{[sSimp]} \quad \frac{A \to B, C \to D}{AC\text{-}B \to D\text{-}(AB)},\ B \cap C \neq \emptyset \neq D \smallsetminus (A \cup B)$$

The thoroughly application of the **[sSimp]** rule to the set Σ_r^* provides an equivalent direct-reduced IS, named Σ_{dr}^*, being the smallest IS fulfilling the following conditions: $\Sigma_r^* \subseteq \Sigma_{dr}^*$ and Σ_{dr}^* is closed under the **[sSimp]** rule.

Once we have got a direct-reduced IS, we can further depurate it by removing extra-attributes and extra-implications thanks to the application of the rules of $\mathbf{SL}_{\text{FD}}$. The target IS for such a depuration step is said to be simplified holding the following conditions: for all $A, B, C, D \subseteq M$,

1. $A \to B$, $A \to C \in \Sigma$ implies $B = C$.
2. $A \to B$, $C \to D \in \Sigma$ and $A \subsetneq C$ imply $C \cap B = \emptyset = D \cap B$.

Thus, for the treatment of Σ^*, the Algorithm *DObasis* has three main stages, each one consisting in the transformation of a previous IS into an equivalent one fulfilling directness and optimality at the end of the process: the direct-optimal basis (see [9] for the proof of this assertion).

In the following, we are going to use this function to transform the first component of a dichotomous IS and compute its equivalent direct-optimal basis.

We begin with the transformation of the original set of implications into a dichotomous one splitting off the treatment for key implications and the others, which provides a better performance of the basis construction method. Note that in Duquenne-Guigues basis key implications are those $A \to B$ that satisfy $A \cup B = M$. These implications have to belong to the second component of the dichotomous set of implications.

Algorithm 1 below structures the above transformation in two consecutive stages: the splitting process (discerning what implications are keys) and then, the direct-optimal transformation for the first component. The following example illustrates the execution of Algorithm 1.

Example 1. Let $\Sigma = \{a \to d, ce \to g, cg \to e, de \to g, bg \to acde, cd \to abeg, abd \to ceg, adeg \to bc\}$ be an IS. In the first stage, we separate the key implications with a linear cost rendering: $\langle \Sigma^*, \Sigma^k \rangle = \langle \{a \to d, ce \to g, cg \to e, de \to g\}, \{bg \to acde, cd \to abeg, abd \to ceg, adeg \to bc\} \rangle$.

In the second stage, we apply Function DObasis to get a direct-optimal basis equivalent to Σ^*: $\Sigma_{do}^* = \{a \to d, ae \to g, ce \to g, cg \to e, de \to g\}$.

Finally, we joint both components of the dichotomous set of implications to get a DD-basis: $\Sigma_{DD} = \langle \{a \to d, ae \to g, ce \to g, cg \to e, de \to g\}, \{bg \to acde, cd \to abeg, abd \to ceg, adeg \to bc\} \rangle$.

6 Conclusions and future works

In this work, we have presented a new definition for ISs in which we characterize two sets of implications with specific properties. In this way, a new direct

Algorithm 1: DD-basis

```
input  : A Duquenne-Guigues basis $\Sigma_{DG}$ on $M$
output: The DD-basis $\Sigma_{DD}$ on $M$
begin
    /* Stage 1: Generation of $\Sigma^*$ and $\Sigma^k$ by disjointing of $\Sigma_{DG}$ */
    $\Sigma^* = \emptyset, \Sigma^k = \emptyset$
    foreach $A \to_{\Sigma_{DG}} B$ do
        if $A \cup B = M$ then add $A \to B$ to $\Sigma^k$;
        else add $A \to B$ to $\Sigma^*$;
    /* Stage 2: Generation of $\Sigma^*_{do}$ by executing the above function of $\Sigma^*$ */
    $\Sigma^*_{do} := DObasis(\Sigma^*)$
    /* Output preparation: Generation of $\Sigma_{dd}$ by jointing $\Sigma^*_{do}$ and $\Sigma^k$ */
    $\Sigma_{DD} := \langle \Sigma^*_{do}, \Sigma^k \rangle$
    return $\Sigma_{DD}$
```

basis and the algorithm that renders this new basis, called DD-basis, have been proposed. The main goal we have achieved is the reduction of the cost of computing a direct basis focusing in one subset of the IS: the first component of the dichotomous set. As future work we are going to extend the proposed algorithm when the input was any IS and make a comparative among algorithms related to directness property.

Acknowledgment

Supported by grant TIN11-28084 and TIN2014-59471-P of the Science and Innovation Ministry of Spain.

References

1. K. V. Adaricheva and J. B. Nation and R. Rand, *Ordered direct implicational basis of a finite closure system*, International Symposium on Artificial Intelligence and Mathematics, ISAIM 2012.
2. W W. Armstrong, *Dependency structures of data base relationships*, Proc. IFIP Congress. North Holland, Amsterdam: 580–583, 1974.
3. K. Bertet, M. Nebut, *Efficient algorithms on the Moore family associated to an IS*, DMTCS, 6(2): 315–338, 2004.
4. K. Bertet, B. Monjardet, *The multiple facets of the canonical direct unit implicational basis*, Theor. Comput. Sci., 411(22-24): 2155–2166, 2010.
5. P Cordero, A. Mora, M. Enciso, I.Pérez de Guzmán, *SLFD Logic: Elimination of Data Redundancy in Knowledge Representation*, LNCS, 2527: 141–150, 2002.
6. B. Ganter, *Two basic algorithms in concept analysis*, Technische Hochschule, Darmstadt, 1984.
7. J.L. Guigues and V. Duquenne, *Familles minimales d'implications informatives résultant d'un tableau de données binaires*, Math. Sci. Humaines: 95, 5–18, 1986.
8. A. Mora, M. Enciso, P. Cordero, and I. Fortes, *Closure via functional dependence simplification*, IJCM, 89(4): 510–526, 2012.
9. E. Rodríguez Lorenzo, K. Bertet, P. Cordero, M. Enciso, A. Mora, *The Direct-optimal Basis via Reductions*, CLA: 145-156, 2014.

A new functional measure of skewness based on the convex transform order

Antonio Arriaza-Gómez, Miguel A. Sordo, and Alfonso Suárez-Llorens

Dpto. Estadística e Investigación Operativa, Universidad de Cádiz,
Facultad de Ciencias, Campus Universitario Río San Pedro s/n,
11510 Puerto Real, Cádiz, Spain
{antoniojesus.arriaza,mangel.sordo,alfonso.suarez}@uca.es

Abstract. The tail behaviour of a probability distribution has been widely studied in order to provide robust tools to deal with risk in different fields, such as financial or insurance risk, best known as actuarial theory. In this paper, a new functional skewness measure from the comparative study of the left and right tails of a distribution is provided. The new measure is based on the convex transform order, which let us compare whenever one distribution has heavier tail than another. We study the properties of the functional measure and we shall prove that it allows to detect a tail property called *symmetry in tails.*

Keywords: skewness, asymmetry, heavy-tailed, convex transform order, risk.

1 Introduction

Asymmetry of a continuous distribution is commonly described as skewness, and it has been widely studied in order to measure meaningful differences in the behaviour of the distribution in respect to some location parameter as the mean, median, mode, etc. It is a general practice to make assertions about the symmetry or asymmetry of a probability density function based on scalar measures. Since most of them use all the information of the distribution to summarize in a single number, they do not capture all the meaning of being a symmetric distribution. There are several scalar measures used to quantify the degree of skewness of a distribution, some of most known are [1], [4], [9] and [11].

Asymmetry of probability distribution has been also studied applying a functional approach. Since the symmetry is a functional concept, it seems suitable to describe asymmetry using asymmetry functions. A partial list of the most important measures taking a functional approach include [2], [3], [5], [6], [7] and [8].

Since asymmetry is essentially influenced by the tail behavior of density function, this lead us to study the symmetry of a distribution from the comparative study of its tails. The tail of a distribution is the portion of distribution corresponding to large or small values of the random variable and its study is relevant in actuarial theory, insurance risk, financial risk, etc. In these fields, those distributions that tend to assign higher probabilities to larger values are specially important, they are known as heavy-tailed. The weight of a tail is a property that can be interpreted as a relative concept (the F distribution has a heavier tail than another G) or as an absolute concept (if F verifies a certain property then F is classified as heavy-tailed).

In this paper, we propose a new functional measure that let us compare the left and the right tail of a distribution F. This functional is based on the convex transform order defined by Van Zwet in [10] and it measures which tail is heavier than the other.

The convex transform order is closely related to skewness and shape of the tail distribution. It arises from the need to state when a non-negative distribution G is more skewed to the right than another non-negative distribution F. Given two non-negative distributions F and G we say that F is smaller than G in the convex transform order, written $F \leq_c G$ if, and only if $G^{-1}F(x)$ is a convex function in its domain. This means that the G distribution is obtained throughout "stretching" the F distribution and thereby there exist a change of shape between both distributions. This change of shape involves in how the probability is distributed, in this case, the G distribution displace more probability than the F distribution to the right therefore it is accepted that the G distribution is heavier than the F distribution.

The main idea is to define a skewness function for probability distributions from the comparative study of its tails, interpreting them as non-negative variables. Before starting, we need to define the left tail distribution and the right tail distribution associated to a distribution F from a quantile $F^{-1}(u)$. From here to forward we will denote $F^{-1}(u) = x_u$.

2 Definitions and properties of the functional measure of skewness

Definition 1. *Let X be a random variable which follows a distribution F, and let u be a number in $(0, \frac{1}{2})$, then we define the left tail distribution from the quantile x_u of X as*

$$(X - x_u)^- = \begin{cases} x_u - X & X \leq x_u \\ 0 & \textit{other case} \end{cases}$$

and the right tail distribution from the quantile x_{1-u} of X as the variable

$$(X - x_{1-u})^+ = \begin{cases} X - x_{1-u} & X \geq x_{1-u} \\ 0 & \textit{other case} \end{cases}$$

Let X be a random variable with probability distribution F (see Figure 1), and let $L = (X - x_u)^-$ and $R = (X - x_u)^+$ be its left and right tails from the quantile x_u and x_{1-u}, respectively. Both variables are non-negative (see Figure 2) and their cumulative distribution function take the value $1 - u$ when these variables are null.

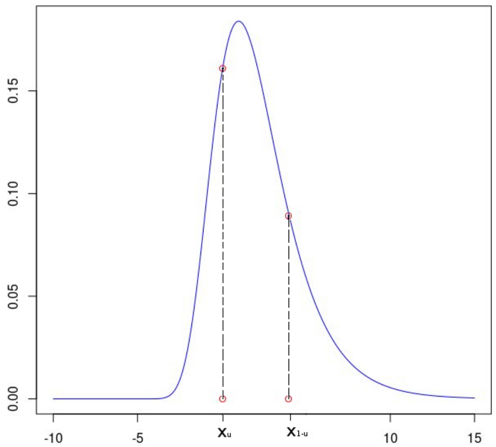


Fig. 1. Probability density function of $X \sim F$

We propose to compare L and R using the convex transform order. However, the convex transform order is a hard condition to be verified, thereby we will use a condition which is implied by the convex transform order. This condition appears in a natural way from a equivalent condition of the convex transform order. We denote F_L and F_R as the probability distribution function of L and R, respectively. If $F_L^{-1}(F_R(x)) \in C^1(\mathbb{R}^+)$, then

$$L \leq_c R \Leftrightarrow (l_v - l_p)f_L(l_v) \geq (r_v - r_p)f_R(r_v) \quad \forall p, v \in (0,1).$$

The above characterization implies the following condition,

$$\int_{1-u}^{1} (l_v - l_{1-u})f_L(l_v)dv \geq \int_{1-u}^{1} (r_v - r_{1-u})f_R(r_v)dv.$$

$$\Leftrightarrow \int_0^u (x_u - x_t)f_X(x_t)dt - \int_{1-u}^{1} (x_v - x_{1-u})f_X(x_v)dv \geq 0.$$

We are now in condition to define a new functional measure of skewness.

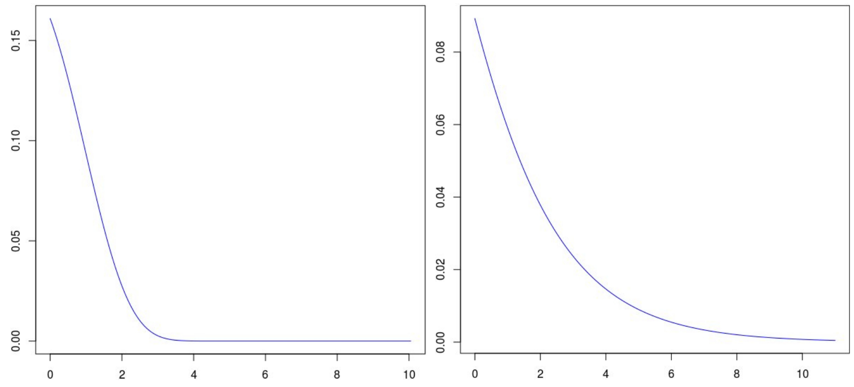


Fig. 2. Plots of probability density function of $(X - x_{1-u})^-$ and $(X - x_{1-u})^-$, respectively

Definition 2. *Let X be a random variable and F its cumulative distribution function, then*

$$S_X(u) = \varphi_X^-(u) - \varphi_X^+(1-u), \quad \forall u \in (0, \frac{1}{2}).$$

where $\varphi_X^-(u) = \int_0^u (x_u - x_v) f_X(x_v) dv$ *and* $\varphi_X^+(1-u) = \int_{1-u}^1 (x_v - x_{1-u}) f_X(x_v) dv$ *are called the left and right skewness measures, respectively.*

Definition 3. *We define the $\mathbb{H}$ set as the set of all continuous random variables X, with probability density function f_X, which verify that:*

1. $\mathbb{E}[f_X(X)] = \int_{-\infty}^{+\infty} f_X^2(x) dx < +\infty.$
2. $\mathbb{E}[X f_X(X)] = \int_{-\infty}^{+\infty} x f_X^2(x) dx < +\infty.$

It has been generally accepted that any measure γ of skewness should satisfy the followings conditions:

1. $\gamma(F) = \gamma(aF + b)$ for all $a > 0$ and all b.
2. $\gamma(F) = -\gamma(-F)$.
3. If $F \leq_c G$, then $\gamma(F) \leq \gamma(G)$. Here, $\leq_c$ denotes the convex transform order.

Proposition 1. *Let $X \in \mathbb{H}$ be a random variable, the functional measure of skewness S_X verifies the three previous conditions.*

Proposition 2. *Let $X \in \mathbb{H}$ be a random variable and f_X its probability density function, then*

If f_X is a decreasing function in its domain $\Longrightarrow S_X(u) \geq 0 \ \ \forall u \in (0, \frac{1}{2})$.
If f_X is an increasing function in its domain $\Longrightarrow S_X(u) \leq 0 \ \ \forall u \in (0, \frac{1}{2})$.
If f_X is a symmetric function $\Longrightarrow S_X(u) = 0 \ \ \forall u \in (0, \frac{1}{2})$.

Corollary 1. *Let $X, Y \in \mathbb{H}$ be two random variable, then*

$$\left.\begin{array}{c}\text{If } X \text{ is symmetric} \\ X \leq_c Y\end{array}\right\} \Rightarrow S_Y(u) \geq 0 \quad \forall u \in (0, \frac{1}{2}).$$

The new skewness measure compares both tails of a distribution from any quantile x_u and its symmetric x_{1-u}. Thereby, it let us detect an interesting property called symmetry in tails, see Figure (3). Also the following result shows that the functional measure of skewness is bounded from a certain value $u_0 \in (0, 1)$ for unimodal distributions.

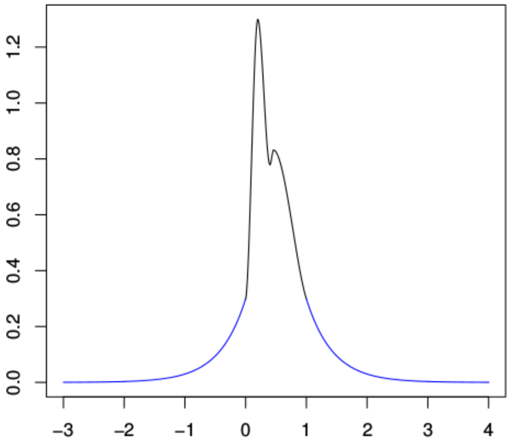


Fig. 3. Probability density function of a tail symmetric distribution

Proposition 3. *Let $X \in \mathbb{H}$ be a random variable with F an strictly increasing function. If X is an unimodal distribution then there exist $u_0 \in (0, \frac{1}{2})$ such that*

$$-\frac{u^2}{2} \leq S_X(u) \leq \frac{u^2}{2} \quad \forall u \leq u_0.$$

References

1. Arnold, B. C. and Groeneveld, R. A. Measuring Skewness with Respect to the Mode. The American Statistician, 49, 34–38 (1995)
2. Benjamini, Y. and Krieger, A.M. Concepts and Measures for Skewness with Data-Analytic Implications. The Canadian Journal of Statistics, La Revue Canadienne de Statistique, 24, 131–140 (1996)
3. Boshnakov, G.N. Some measures for asymmetry of distributions. Statistics & Probability Letters, 77, 1111 - 1116 (2007)
4. Charlier, C.V.L. ber das Fehlergesetz. Arkiv fir Mathematik, Astronomi och Fysik, 2.8. (1920)
5. David, F.N. and Johnson, N.L. Some Tests of Significance with Ordered Variables, J.R. Statist. Soc. Ser. B, 18,1–20 (1956)
6. Doksum, K.A. Measures of Location and Asymmetry, Scand. J. Statist., 2, 11–22 (1975)
7. Groeneveld, R. A. and Meeden, G. Measuring Skewness and Kurtosis, Journal of the Royal Statistical Society. Series D (The Statistician), 33, 391– 399 (1984)
8. MacGillivray, H.L. Skewness and Asymmetry: Measures and Orderings, Ann. Statist., 14, 994–1011 (1986)
9. Pearson, K. Contributions to the Mathematical Theory of Evolutions, II: Skew Variations in Homogeneous Material. Transactions of the Royal Philosophical Society, Ser. A, 186, 343–414 (1985)
10. Van Zwet, W.R. Convex Transformations of Random Variables, in Mathematical Centre Tracts 7, Mathematisch Centrum, Amsterdam (1964)
11. Yule, G.W. An Introduction to the Teory of Statistics (2nd ed.), London: Griffin (1912)

An LLVM-based Approach to Generate Energy Aware Code by means of MOEAs

Sébastien Varrette[1], Bernabé Dorronsorro[2] and Pascal Bouvry[1]

[1] Computer Science and Communication (CSC) Research Unit
University of Luxembourg, 6, rue Richard Coudenhove-Kalergi
L-1359 Luxembourg, Luxembourg
[2] University of Cadiz
Avenida de la Universidad, 10,
11519 Puerto Real, Cadiz, Spain

Abstract. Moderating the energy consumption and building eco-friendly computing infrastructure is of major concerns in the implementation of High Performance Computing (HPC) system, especially when a worldwide effort target the production of an Exaflop machine by 2020 within a power envelop of 20 MW. Tracking energy savings can be done at various levels and in this paper, we investigate the automatic generation of energy aware software with the ambition to keep the same level of efficiency, testability, scalability and security.
To this end, the Evo-LLVM framework is proposed. Based on the modular LLVM Compiler Infrastructure and exploiting various evolutionary heuristics, our scheme is designed to optimize for a given input source code (written in C) the sequence of LLVM transformations that should be applied to the source code to improve its energy efficiency without degrading its other performance attributes (execution time, parallel or distributed scalability). Measuring this capacity is based on the combination of several metrics optimized simultaneously with Multi-Objective Evolutionary Algorithms (MOEAs). In this position paper, the NSGA-II algorithm is implemented within the Evo-LLVM yet the analysis of more advanced heuristics is in progress. In all cases, the experimental validation of the framework over a pedagogical code sample reveal a drastic improvement of the energy consumed during the execution while maintaining (or even improving) the average execution time.

Keywords: Performance evaluation, Energy-efficiency, HPC, Evolutionary Algorithm, Fault-Tolerance Result-Checking

1 Introduction

Energy management has become a key challenge in the area of computing systems today. For large scale systems, such as data centers, energy efficiency has proven to be the key for reducting all kind of costs related to capital, operational expenses and environmental impact. Power drainage of a system is closely related to the type and characteristics of workload that the device is running. These characteristics refer to the way the workload utilizes different resources and components of the system, such as CPU, memory, disc etc. Modern system design now includes components that support energy management at various level, for instance through a dynamic scaling of the power (or frequency) allocated to its usage (DVFS for the CPU etc.) and/or an integrated way to handle idle state for a more or less long period of time. In this paper, we take advantage of these techniques and the corresponding sensors embedded within the Linux kernel to estimate the average power consumption induced by the execution of a given process. Combined with other metrics quantifying the inherent performance of the execution, it is thus possible to design a Multi-Objective Evolutionary Algorithm (MOEA) system able to evolve a given source code (called the *reference* source in the sequel) to produce an set of energy-aware versions able to compete from the pure execution time point of view with the initial performance of the reference source. This idea led to the design of the Evo-LLVM framework presented in this paper.

This article is organized as follows: section 2 details the background of this work and reviews related works. Then, the Evo-LLVM framework is presented in the section 3. Implementation details of the proposed framework are provided in the section 3. The validation of the approach on a concrete benchmarking code is expounded in the section 4 which details and discusses the experimental results obtained. Finally, the section 5 concludes the paper and provides some future directions and perspectives opened by this study.

2 Context & Motivations

Since the advent of high-level programming language, research in the compilation domain have always seek to automate and find novel optimization techniques to produce a compiled code that improve the running time. In this context, many previous studies identified a large number of transformations that could be applied to the different section of a source code to generate different and hopefully improved version of the compiled executable. A reference summary of these transformations, their effects and their respective application context is described in [1]. Determining the optimal sequence of transformations to apply to a given source code that would minimize the execution time over a given computing system is proven to be an NP-complet [9] problem. It follows that all modern compilers such as GCC (the the GNU Compiler Collection) or LLVM rely on static heuristics involving a subset of transformations applied in an order that grant, in general, good results n general while ensuring a bounded compilation time [10]. Because of all these factors, the optimization operated by compilers hardly produces "optimal" output in any sense, and may even impede performances in some cases. It follows a considerable optimization work so as to try a set of transformation potentially valuable. This time-consuming process is generally performed by hand and requires expert engineering skills. The current state-of-the-art tries to address optimization problem from a transverse way, *i.e.* by means of automatic analysis schemes generally based on Evolutionary heuristics. For instance, a genetic approach is done in [2] to optimize the size of the output binaries. Also, the Acovea [8] framework (Analysis of Compiler Options via Evolutionary Algorithm) for `gcc` or Cole [7] investigates in an automatic manner the best combination of compiler options leading to the fastest executable program from a given source code. Complementary, recent advances over a new kind of software development environment inspired from Search Based Software Engineering (SBSE) [6] led to the definition of GISMOE challenge (Genetic Improvement of Software for Multiple Objective Exploration) [5]. The general idea is that it is possible to combine the recent advances in software test data generation, genetic programming and multi objective optimization to build a development environment capable of producing a Pareto program surface that would help the software designer to navigate between different version of the same program (typically the execution time, the memory usage and the energy efficiency). The work proposed in this paper definitively offers the basic building block able to propose a concrete answer to this challenge.

2.1 The LLVM Compiler Infrastructure

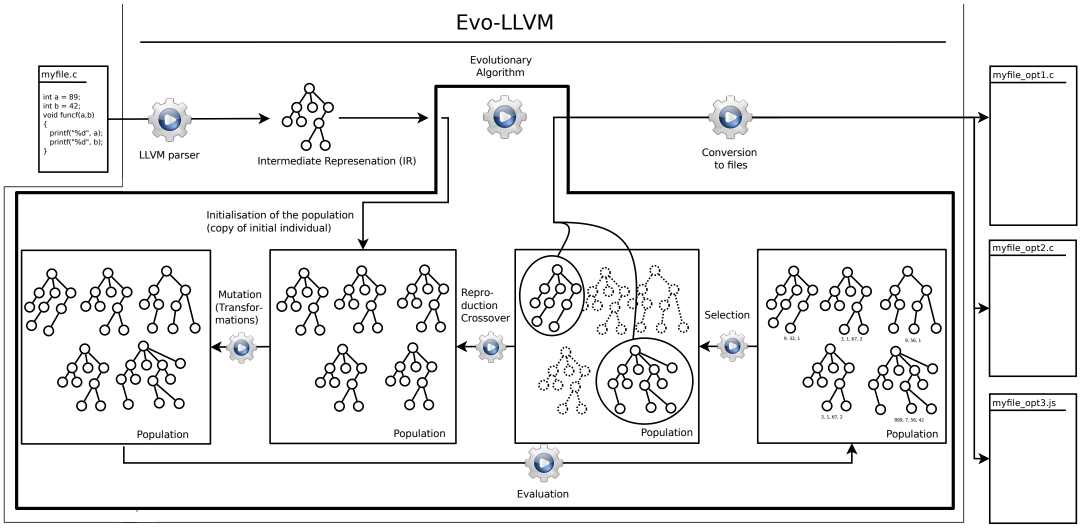
The LLVM compiler infrastructure project (formerly Low Level Virtual Machine) is a compiler infrastructure designed to be a set of reusable libraries with well-defined interfaces. It is written in C++ and is designed for compile-time, link-time, run-time, and "idle-time" optimization of programs written in arbitrary programming languages. LLVM was originally written to be a replacement for the existing code generator in the GCC stack and many of the GCC front ends have been modified to work with it. Widespread interest in LLVM has led to a number of efforts to develop entirely new front ends for a variety of languages. The one that has received the most attention is Clang, a new compiler supporting C, Objective-C and C++ supported by Apple. The core of LLVM is the intermediate representation (IR), a low-level programming language similar to assembly. IR is a strongly typed RISC instruction set which abstracts away details of the target.

In this article, we propose to exploit the flexibility offered by LLVM to manipulate the IR modelization of a given source code to check the opportunity of applying a sequence of supported transformations and evaluating the impact on the energy efficiency of the produced executable. The choice of the transformation to apply and their order shall be governed by an evolutionary heuristic. The validation of the approach shall be performed on a relevant set of benchmark applications.

2.2 Evolutionary Algorithms (EAs)

EA is a class of solving techniques based on the Darwinian theory of evolution [3] which involves the search of a *population* X_t of solutions. Members of the population are feasible solutions and called *individuals*. Each iteration of an EA involves a competitive selection that weeds out poor solutions through the *evaluation* of a fitness value that indicates the quality of the individual as a solution to the problem. The evolutionary process involves at each generation a set of stochastic operators that are applied on the individuals, typically recombination (or cross-over) and mutation.

3 The Evo-LLVM Compiler Framework



```
int a = 89;
int b = 42;
void funcf(a,b)
{
   printf("%d", a);
   printf("%d", b);
}
```

Fig. 1. Overview of Evo-LLVM framework, describing the full process of the generation of new representations of the code.

This section briefly review the Evo-LLVM framework as a natural extension of the Shadobf framework proposed in [**?**]. The general code optimization process operated by Evo-LLVM is illustrated in the figure 1: from the initial program $\mathcal{P}$ to be analysed (`myfile.c` in the figure), a reference individual (in the EA sense) I_{ref} is generated that represents $\mathcal{P}$. Then, a complete population of n individuals is generated by randomly applying a mutation (*i.e.* an LLVM transformation) on the reference individual I_{ref}. The MOEA process (NSGA-II in the current state of the implementation) then intervene to explore the search space induced by the different objectives seek for the produced binaries, evaluate the population and apply the genetic operators (mutation and cross-over). This permits to exhibit at each generation non-dominated Pareto solutions, each of them representing a set of derived (and hopefully more performant for all considered metrics) versions of the program $\mathcal{P}$ that propose a good trade-off between each objectives *i.e.* metrics.

The key characteristics of the Evo-LLVM framework are as follows:

- The C code is parsed using LLVM to produced the intermediate representation of the program (IR).

- LLVM has 54 built-in transformations. These range from tail call elimination (a method to optimize some recursive functions) to loop unwinding (reducing loop overhead). The order these transforms are applied can matter: for example, dead instruction elimination might not find an unused instruction in an unoptimized program, but after a few passes of other transforms, some instructions may be superfluous. This is important to keep in mind when designing the evolutionary algorithm, specifically when deciding on the crossover methods. Some of these may split up two transformations that only work well in tandem. In all cases, the LLVM transforms are randomly applied within Evo-LLVM individuals during the evolutionary operators;
- Throughout NSGA-II [4] (one of the reference selection algorithm for MOEAs considered in our initial implementation), *individuals* are selected by taking into account the non-domination criteria and the distance from one to the others to guarantee a good diversity as well as the leading *individuals* of the population. The concept of dominance is the following (in the case of minimization): an *individual* I with the objectives values $f_{obj}(I)$ is said to be dominated by J if

  $$\forall obj \in objectives, f_{obj}(J) < f_{obj}(I)$$

  An *individual* is said to be non-dominated if it is not dominated by any other *individuals* in the population. All the non dominated solutions of a population are the approximated Pareto front of the problem. NSGA-II is selecting all the non-dominated solutions of the population, and if the size of the new population is lower than the maximum size, NSGA-II is selecting again all the non-dominated solutions of the old population but this time excluding the already selected *Individuals.*
- The performance metrics permits to evaluate each individuals.
- The IR model might be helpful to compute additional static metrics (sequential work, number of instructions).

4 Validation and Experimental Results

We have validated our approach a simple pedagogical example *i.e.* a quicksort algorithm. It was chosen as it involves many sections in the code that are worth optimizing: memory allocation, iterations, recursion and branching, all intertwined. However, the program is independent from the chosen algorithm, a few changes in the configuration make it possible to run the optimization on any algorithm.

For the sake of simplicity, we show here on short runs involving the simultaneous evolution of the power consumption along with the execution time when the number of benchmarks per individual is set to 100. A brief overview of the generated Pareto front is proposed.

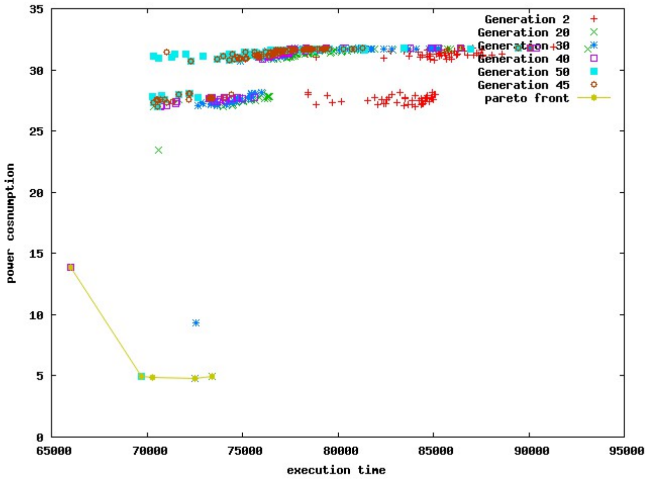


Fig. 2. Set of the 2D Pareto fronts approximation for the quicksort program using NSGA-II and Evo-LLVM

The table 1 shows the characteristics of a set of individuals selected after application of Evo-LLVM. In practice, we selected the individual which is the closest to the median values of every individuals which are in the Pareto front, allowing to have a good trade off between all the objectives. In all cases we see that the selected individuals upon successive generations demonstrate an interesting improvement in terms of power consumption while not degrading the execution time.

Generation	Power	Execution time
$2^{nd}_{reference}$	31.0262	70554.0
20^{th}_{best}	9.3593	72573.0
45^{th}_{best}	4.9078	70266.0
50^{th}_{best}	4.9183	69694.0

Table 1. Energy and execution time comparison for generation 20,45,50 for selected individuals after Evo-LLVM run.

5 Conclusion

The main objective of this work was to proceed to the automatic generation of energy aware software while maintaining the same level of efficiency, testability and scalability. As it is well-known, power drainage of a system is not a static property that depends solely on hardware characteristics. Using energy aware software will lead to significant reduction to the overall energy consumption. The benefits will reach not only large scale computing systems or data canters but also home users that aim to a more efficient energy management. To this end, the Evo-LLVM framework is proposed. Based on the modular LLVM Compiler Infrastructure and exploiting various evolutionary heuristics, our scheme is designed to optimize for a given input source code (written in C) the sequence of LLVM transformations that should be applied to the source code to improve its energy efficiency without degrading its other performance attributes (execution time, parallel or distributed scalability). Measuring this capacity is based on the combination of several metrics optimized simultaneously with Multi-Objective Evolutionary Algorithms (MOEAs). In this position paper, the NSGA-II algorithm is implemented within the Evo-LLVM yet the analysis of more advanced heuristics is in progress. Experimental results on a simple pedagogical program demonstrated an 84.20% improvement on the average consumed energy while not degrading the execution time.

Acknowledgments: The experiments presented in this paper were carried out using the HPC facility of the University of Luxembourg.

References

1. D. F. Bacon, S. L. Graham, and O. J. Sharp. Compiler transformations for high-performance computing. *ACM Comput. Surv.*, 26(4), 1994.
2. K. D. Cooper, P. J. Schielke, and D. Subramanian. Optimizing for reduced code space using genetic algorithms. In *Workshop on Languages, Compilers, and Tools for Embedded Systems*, pages 1–9, 1999.
3. C. Darwin. *On the Origin of Species by Means of Natural Selection*. Londres, 1859.
4. K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: Nsga-ii. pages 849–858. Springer, 2000.
5. M. Harman, W. B. Langdon, Y. Jia, D. R. White, A. Arcuri, and J. A. Clark. The gismoe challenge: Constructing the pareto program surface using genetic programming to find better programs (keynote paper). In *Proceedings of the 27th IEEE/ACM International Conference on Automated Software Engineering*, ASE 2012, pages 1–14, New York, NY, USA, 2012. ACM.
6. M. Harman, P. McMinn, J. de Souza, and S. Yoo. Search based software engineering: Techniques, taxonomy, tutorial. In B. Meyer and M. Nordio, editors, *Empirical Software Engineering and Verification*, volume 7007 of *Lecture Notes in Computer Science*, pages 1–59. Springer Berlin Heidelberg, 2012.
7. K. Hoste and L. Eeckhout. Cole: compiler optimization level exploration. In *CGO*, pages 165–174, 2008.
8. S. Ladd. Acovea: Using Natural Selection to Investigate Software Complexities. [Online] see `www.coyotegulch.com/products/acovea/`, 2007.
9. A. Nisbet. GAPS: A Compiler Framework for Genetic Algorithm (GA) Optimised Parallelisation. In *HPCN Europe*, pages 987–989, 1998.
10. R. Stallman et al. *Using GCC: The GNU Compiler Collection Reference Manual*. FSF, 2005.

Detecting User Influence in Twitter: PageRank vs Katz, a case study*

Hugo Rosa[1], Joao P. Carvalho[1,2], Ramon Astudillo[1], and Fernando Batista[1,3]

[1] INESC-ID
[2] Instituto Superior Técnico, Universidade de Lisboa
[3] ISCTE-IUL - Instituto Universitário de Lisboa
{hugo.rosa,joao.carvalho,ramon.astudillo,fmmb}@inesc-id.pt

Abstract. Microblogs, such as Twitter, have become an important socio-political analysis tool. One of the most important tasks in such analysis is the detection of relevant actors within a given topic through data mining, i.e., identifying who are the most influential participants discussing the topic. Even if there is no gold standard for such task, the adequacy of graph based centrality tools such as PageRank and Katz is well documented. In this paper, we present a case study based on a "London Riots" Twitter database, where we show that Katz is not as adequate for the task of important actors detection since it fails to detect what we refer to as "indirect gloating", the situation where an actor capitalizes on other actors referring to him.

Keywords: Page Rank, Katz, User Influence, Twitter, Data Mining

1 Introduction

Nowadays, there are 288 million active users on Twitter and more than 500 million tweets are produced per day [16]. The impact of Twitter on the Arab Spring [5] and how it beat the all news media to the announcement of Michael Jackson's death [14], are just a few examples of Twitter's role in society. When big events occur, it is common for users to post about it in such fashion, that it becomes a trending topic, all the while being unaware from where it stemmed or who made it relevant. The question we wish to answer is: "Which users were important in disseminating and discussing a given topic?".

Determining user relevance is vital to help determine trend setters [15]. The user's relevance must take into account not only global metrics that include the user's level of activity within the social network, but also his impact in a given topic [17]. Empirically speaking, an influential person can be described as someone with the ability to change the opinion of many, in order to reflect his own.

* This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) under project PTDC/IVC-ESCT/4919/2012 and funds with reference UID/CEC/50021/2013.

While [12] supports this statement, claiming that “a minority of users, called influentials, excel in persuading others”, more modern approaches [4] seem to emphasize the importance of interpersonal relationships amongst ordinary users, reinforcing that people make choices based on the opinions of their peers. In [2], three measures of influence were taken into account: “in-degree is the number of people who follow a user; re-tweets mean the number of times others forward a user’s tweet; and mentions mean the number of times others mention a user’s name.”. It concluded that while in-degree measure is useful to identify users who get a lot of attention, it “is not related to other important notions of influence such as engaging audience”. Instead “it is more influential to have an active audience who re-tweets or mentions the user”. In [7], the conclusion was made that within Twitter, “news outlets, regardless of follower count, influence large amounts of followers to republish their content to other users”, while “celebrities with higher follower totals foster more conversation than provide retweetable content”. The authors in [11] created a framework named “InfluenceTracker”, that rates the impact of a Twitter account taking into consideration an Influence Metric, based on the ratio between the number of followers of a user and the users it follows, and the amount of recent activity of a given account. Much like [2], it also shows that “that the number of followers a user has, is not sufficient to guarantee the maximum diffusion of information (...) because, these followers should not only be active Twitter users, but also have impact on the network”.

With the previous definitions of influence in mind, we propose a graph representation of user’s influence based on “mentions”. Whenever a user is mentioned in a tweet’s text, using the @*user* tag, a link is made from the creator of the tweet, to the mentioned user, regardless of it being a retweet or a conversation. For example, the tweet ”*Do you think we can we get out of this financial crisis, @userB?*”, from @*userA*, creates the link: @*userA* $\longrightarrow$ @*userB*.

2 Network Analysis Algorithms

In graph theory and network analysis, the concept of centrality refers to the identification of the most important vertices’s within a graph, i.e., most important users. We therefore define a graph $G(V, E)$ where V is the set of users and E is the set of directed links between them. Arguably the most well known centrality algorithm is PageRank [8]. It is one of Google’s methods to its search engine and uses web pages as nodes, while back-links form the edges of the graph. It is defined by Equation 1 as $PR(v_i)$ of a page v_i.

$$PR_{vi} = \frac{1-d}{N} + d \sum_{v_j \in M(v_i)} \frac{PR(v_j)}{L(v_j)} \tag{1}$$

In Equation 1, v_j is the sum ranges over all pages that has a link to v_i, $L(v_j)$ is the number of outgoing links from v_j, N is the number of documents/nodes in the collection and d is the damping factor. The PageRank is considered to be a random walk model, because the weight of a page v_i is ”the probability that a random walker (which continues to follow arbitrary links to move from page to page) will be at v_i at any given time. The damping factor corresponds to the probability of the random walk to jump to an arbitrary page, rather than to follow a link, on the Web. It is required to reduce the effects on the PageRank computation of loops and dangling links in the Web." [10]. The true value that Google uses for damping factor is unknown, but it has become common to use $d = 0.85$ in the literature. A lower value of d implies that the graph's structure is less respected, therefore making the "walker" more random and less strict.

Another well known method is the Katz algorithm [6]. It is a generalization of a back-link counting method where the weight of each node is "determined by the number of directed paths that ends in the page, where the influence of longer paths is attenuated by a decay factor" and "the length of a path is defined to be the number of edges it contains" [10]. It is defined by Equation 2 "where $N(v_i, k)$ is the number of paths of length k that starts at any page and ends at v_i and α is the decay factor. Solutions for all the pages are guaranteed to exist as long as α is smaller than $\lambda > 1$, where $1/\lambda$ is the maximum in-degree of any page" [10].

$$I_{vi} = \sum_{k=0}^{\infty}[\alpha^k N(v_i, k)] \tag{2}$$

3 Experiments and Results

In order to test the network analysis methods presented above, a database from the London Riots in 2011 [3] was used. The Guardian Newspaper made public a list of tweets from 200 influential twitter users, which contains 17795 riot related tweets and an overall dataset of 1132938 tweets. Using a Topic Detection algorithm [1], we obtained an additional 25757 unhastagged tweets about the London Riots. It consists of a Twitter Topic Fuzzy Fingerprint algorithm [13] that provides a weighted rank of keywords for each topic in order to identify a smaller subset of tweets within scope. The sum of posting and mentioned users is 13765 (vertices) and it has 19993 different user mentions (edges), achieving a network connectivity ratio of $\frac{edges}{vertices} = 1.46$.

The remainder of this section presents the results of each algorithm's ranking for most influential users. An empirical study of the users is made, in order to ascertain their degree of influence. The graphs and ranking were calculated using *Graph-Tool* [9].

Table 1 shows how both network analysis algorithms behave with our graph representation, while highlighting the changes in rank between them, as shown by the arrows in the last column. Figure 1 provides a visual tool to the graph, as provided by PageRank. There is a relation between the number of mentions and the ranking in both algorithms, since these users are some of the most mentioned users in our dataset.

When comparing PageRank with Katz, several differences arise, but the top two users are agreed upon: i) @guardian, Twitter account of the world famous

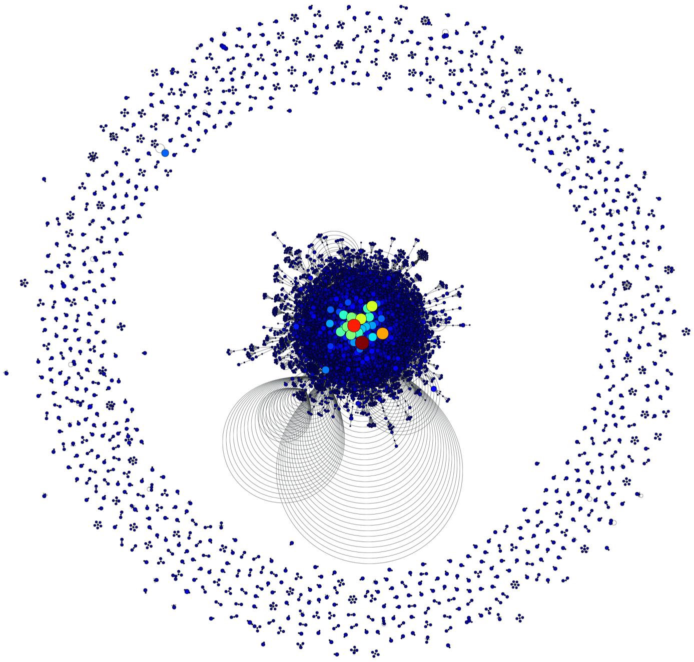


Fig. 1. User influence Page Rank Graph - larger circles indicate larger user influence.

Table 1. Most influential users according to Page Rank, and comparison with Katz.

User	Mentions		PageRank		Katz		
	#	rank	score	rank	score	rank	
@guardian	160	2	0.0002854	1	0.022157	2	
@skynewsbreak	178	1	0.0002512	2	0.023479	1	
@gmpolice	122	4	0.0002128	3	0.019009	4	
@riotcleanup	107	6	0.0001767	4	0.017992	6	↘
@prodnose	67	14	0.0001761	5	0.014022	15	↘↘↘
@metpoliceuk	116	5	0.0001494	6	0.018709	5	
@marcreeves	69	11	0.0001476	7	0.014195	12	↘↘
@piersmorgan	78	8	0.0001465	8	0.014959	9	
@scdsoundsystem	69	12	0.0001442	9	0.014190	13	↘↘
@subedited	70	10	0.0001337	10	0.014278	11	
@youtube	48	20	0.0001257	11	0.012424	20	↘↘↘
@bbcnews	94	7	0.0001256	12	0.016426	8	↗↗
@mattkmoore	62	15	0.0001237	13	0.013614	16	↘
...							
@paullewis	129	3	0.0000954	20	0.019602	3	↗↗↗↗
...							
@juliangbell	61	16	0.0000275	188	0.0166597	7	↗↗↗↗↗↗↗

newspaper “The Guardian”; ii) @skynewsbreak, Twitter account of the news team at Sky News TV channel. This outcome agrees with [7] previous statement, that, “news outlets, regardless of follower count, influence large amounts of followers to republish their content to other users”. Other users seem to fit the profile, namely @gmpoliceq and @bbcnews. Most of the other users are either political figures, political commentators or jornalists (@marcreeves, @piersmorgan, and @mattkmoore).

However, Katz’s third and seventh top ranked users, are not in PageRank’s top users. These are two very different cases: i) @paullewis, ranked 3rd by Katz shows up at 20th according to PageRank; ii) @juliangbell, ranked 7th by Katz shows up at 188th according to PageRank. The reason behind @paullewis high placement in the Katz rank is the number of mentions. As said previously, Katz is a generalization of a back-link counting method, which means the more back-links/mentions a user has, the higher it will be on the ranking. This user has 129 mentions, but PageRank penalizes it, because it is mentioned by least important users, which means a less sum weight is being transfered to it in the iterative process. This logic also applies to user @bbcnews. Additionally, @paullewis is also an active mentioning user, having mentioned other users a total of 14 tweets, while @skynewsbreak and @guardian have mentioned none. As a consequence, Paul Lewis transfers its influence across the network while the others simply harvest it. There are several users that drop in ranking from PageRank to Katz for the very same reason. Users such as @prodnose, @marcreeves and @youtube do not have enough mentions for Katz to rank them higher. User @juliangbell, despite mentioned often (61 times), is down on the PageRank because of indirect gloating, i.e., he retweets tweets that are mentioning himself: *“@LabourLocalGov #Ealing Riot Mtg: @juliangbell speech http://t.co/3BNW0q6”* was posted by @juliangbell himself. The user is posting somebody else’s re-tweet of one of his tweets. As a consequence a link/edge was created from @juliangbell to @LabourLocalGov, but also from @juliangbell to himself, since his username is mentioned in his own tweet. Julian Bell is a political figure, making it acceptable that he would have a role in discussing the London Riots, but the self congratulatory behavior of re-tweeting other people’s mentions of himself, is contradictory with the idea of disseminating the topic across the network. While Katz is not able to detect this effect, PageRank automatically corrects it. Contrary to what is mentioned in previous works, it is our comprehension that Katz is not adequate to detect a user’s importance in social media such as Twitter.

4 Conclusions and Future Work

With this study, we have shown that in the context of user influence in Twitter, PageRank and Katz are not equal in performance, thus disproving previous claims. PageRank has proved a more robust solution to identify influential users in discussing and spreading a given relevant topic, specially when considering how it deals with indirect gloating, an item Katz fails to penalize.

References

1. Carvalho, J.P., Pedro, V., Batista, F.: Towards intelligent mining of public social networks' influence in society. In: IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS). pp. 478 – 483. Edmonton, Canada (June 2013)
2. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.P.: Measuring user influence in twitter: The million follower fallacy. In: in ICWSM '10: Proceedings of international AAAI Conference on Weblogs and Social (2010)
3. Crockett, K, S.R.: Twitter riot dataset (tw-short) (2011)
4. Domingos, P., Richardson, M.: Mining the network value of customers. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 57–66. KDD '01, ACM, New York, NY, USA (2001), http://doi.acm.org/10.1145/502512.502525
5. Huang, C.: Facebook and twitter key to arab spring uprisings: report. http://www.thenational.ae/news/uae-news/facebook-and-twitter-key-to-arab-spring-uprisings-report (June 2011), accessed: 2014-05-02
6. Katz, L.: A new status index derived from sociometric analysis. Psychometrika 18(1), 39–43 (March 1953), http://ideas.repec.org/a/spr/psycho/v18y1953i1p39-43.html
7. Leavitt, A., Burchard, E., Fisher, D., Gilbert, S.: The influentials: New approaches for analyzing influence on twitter (2009)
8. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web (1999)
9. Peixoto, T.: https://about.twitter.com/company
10. Phuoc, N.Q., Kim, S.R., Lee, H.K., Kim, H.: Pagerank vs. katz status index, a theoretical approach. In: Proceedings of the 2009 Fourth International Conference on Computer Sciences and Convergence Information Technology. pp. 1276–1279. ICCIT '09, IEEE Computer Society, Washington, DC, USA (2009), http://dx.doi.org/10.1109/ICCIT.2009.272
11. Razis, G., Anagnostopoulos, I.: Influencetracker: Rating the impact of a twitter account. CoRR (2014), http://arxiv.org/abs/1404.5239
12. Rogers, E.M.: Diffusion of innovations (1962)
13. Rosa, H., Batista, F., Carvalho, J.P.: Twitter topic fuzzy fingerprints. In: WCCI2014, FUZZ-IEEE, 2014 IEEE World Congress on Computational Intelligence, International Conference on Fuzzy Systems. pp. 776–783. IEEE Xplorer, Beijing, China (July 2014)
14. Sankaranarayanan, J., Samet, H., Teitler, B.E., Lieberman, M.D., Sperling, J.: Twitterstand: News in tweets. In: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. pp. 42–51. GIS '09, ACM, New York, NY, USA (2009), http://doi.acm.org/10.1145/1653771.1653781
15. Tinati, R., Carr, L., Hall, W., Bentwood, J.: Identifying communicator roles in twitter. In: Proceedings of the 21st International Conference Companion on World Wide Web. pp. 1161–1168. WWW '12 Companion, ACM, New York, NY, USA (2012), http://doi.acm.org/10.1145/2187980.2188256
16. Twitter: https://about.twitter.com/company
17. Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: Finding topic-sensitive influential twitterers. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining. pp. 261–270. WSDM '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1718487.1718520

Simplification of Inference Problems Based on High Dimensional Vectors by Wavelet Transformation and Fuzzy Rule Interpolation

Ferenc Lilik[1], Szilvia Nagy[1], and László T. Kóczy[1,2]

[1] Széchenyi István University
H-9026 Győr, Hungary,
lilikf@sze.hu
[2] Budapest University of Technology and Economics
H-1117 Budapest, Hungary

Abstract. A new approach for inference based on treating sampled functions is presented. Sampled functions can be transformed into only a few points by wavelet analysis, thus the complete function is represented by these several discrete points. The finiteness of the teaching samples and the resulting sparse rule bases can be handled by fuzzy rule interpolation methods, like, e.g., KH interpolation. Using SHDSL transmission performance prediction as an example, the simplification of inference problems based on large, sampled vectors by wavelet transformation and fuzzy rule interpolation applied on these vectors are introduced in this paper.

Keywords: Fuzzy inference, performance prediction, fuzzy rule interpolation, wavelet analysis, telecommunications access networks

1 Introduction

Due to the great number of input values, making inference on phenomena which can be described by large-sized vectors are difficult and expensive. In order to construct efficient inference systems, simplification of the input space is needed. This simplification makes the process of the inference easier, however, it unavoidably rises the system's level of uncertainty and inaccuracy. During our previous research on performance prediction of physical links of telecommunications access networks, we had to encounter such problems in two ways. Horizontally, making decisions by the observation of only a part of the physical reality resulted in sparse fuzzy rule bases. Vertically, drastically lowering the number of the measured frequency dependent input values caused an inaccuracy in the final results.

In Section 2 the primary technical problem underlying the research on performance prediction is briefly reviewed. In Section 3 wavelet transformation and fuzzy rule interpolation as the algorithmic techniques applied in a combined way for handling the problems of simplification are outlined, and in Section 4 we present the test results of the new approach based on these techniques.

2 SHDSL Performance Prediction of physical links in telecommunications access networks

In our previous work [1, 2] we have laid the foundations of a Mamdani-type fuzzy inference method for pre-qualification of telecommunication access network links based on measured insertion loss and noise values of the given lines. We applied fuzzy rule bases of two types, the one was generated from the measured data's statistical properties using triangular sets, the other was generated by an evolutionary algorithm using trapezoidal sets [3]. Examples of the resulting rules can be seen in Fig. 1.

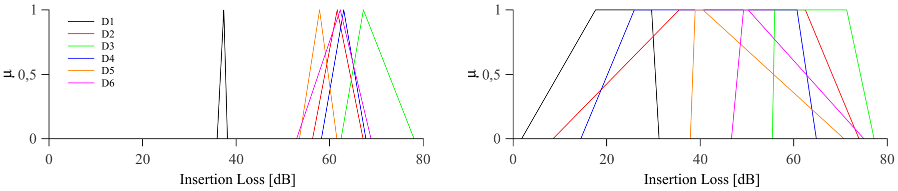


Fig. 1. Examples of rule antecedents from our previous predicting methods [2].

The above two rule bases were tested by the measurements of more than 60 wire pairs in operating access networks and there were no relevant differences between their respective results. In most of the cases, where all measured values belonged to insertion loss areas covered by antecedent sets, the predictions were successful. Only 13 lines out of 65 could be evaluated, and the predictions were correct in case of 12 lines form these 13.

3 Methods for handling the vertical and horizontal sparseness

The reason for the insufficient performance of the pre-qualification method is the two-dimensional sparseness of the inference system.

Vertical sparseness of the rule bases was derived from the partial usage of the possible input data. It was needed in order to decrease the dimensionality of the applied fuzzy inference system, however, a lot amount of information of the measured insertion loss functions was wasted. Finding a method which keeps the simplicity of the fuzzy system and the information of the used insertion loss functions was needed. As wavelet transformation is efficient in reducing the size of any continuous or discrete functions down to a required level, it seemed to be successfully applicable in the problem.

Horizontal sparseness of the fuzzy system, namely the sparseness of the rule bases, can be handled by the techniques of fuzzy rule interpolation. Stabilized KH interpolation fits continuous and mathematically stable functions to all α-cuts of the membership functions in the rules, which can tackle the observations

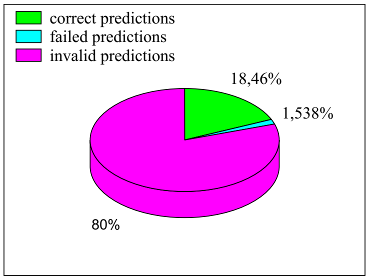


Fig. 2. Success rate of the rule bases.

in the gaps and out of the domains of the rules too (in this way performing also extrapolation).

Basics of wavelet transformation and stabilized KH interpolation are briefly overviewed in the followings.

3.1 On wavelet analysis

In data processing in general wavelet theory [5] has proved to be a very useful tool. The largest part of the methods use wavelets is the image compression [6] and data analysis, but it can also be used for solving differential equations [7].

Wavelet transform and of a function provides data about the function's fine-scale and rough-scale behavior. Wavelet analysis can be carried out by a series of filter pairs. There is a high-pass and a low-pass filter in all of the pairs, the high-pass ones (after a downsampling) giving the wavelet (detail) components and the low-pass ones being transformed further, as it can be seen in Fig. 3.

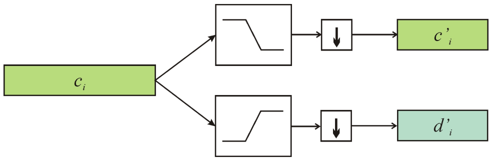


Fig. 3. One filter pair of the discrete wavelet transform. After the high pass and low pass convolutional filters and the downsamplings the transformed vectors c_i' and d_i' arise, their size is about half of the size of the original c_i.

In data analysis – also in our case – the starting point is a sampled function and the end result is the lowest resolution level low pass vector and the high pass vectors. Our starting vector is a series of insertion loss values measured at consecutive frequency points, and the resulting vectors give information about the large-scale behavior of the insertion loss vs. frequency function. In the following considerations Haar's [8] and Daubechies's [5] wavelet and scaling function sets are used with 2 and 4 nonzero filter coefficients, respectively. Transformations of the starting sampled insertion loss functions were carried out until only 5 vector elements remained.

3.2 Stabilized KH rule interpolation

In case of sparse rule bases, KH interpolation [9, 10] is a mathematically stable and widely applicable fuzzy rule interpolation method. Its improved version is the stabilized KH interpolation. In our work we used this improved technique in order to eliminate the problems originating from the sparseness of the rule bases.

The method is based on the distances between the examination vector and the antecedent sets of the rule base. The closures of the α-cuts of the interpolated resolution are given in [11].

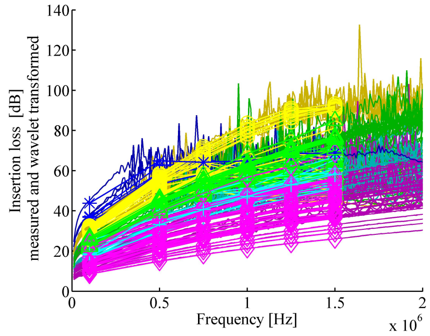


Fig. 4. Insertion loss values and the corresponding wavelet transforms. Different performance classes are indicated by different colors.

4 A new prediction method based on the combination of wavelet transformation and stabilized KH rule interpolation

In order to avoid the problems reviewed in Section 1, the techniques of Section 3.1 and 3.2 were used.

First, the wavelet transformed version of the insertion loss values used in rule base construction were calculated. Daubechies-2 (Haar) [8] and Daubechies-4 wavelets were used and the transformations were performed down to 5 points resolution. Fig. 4 shows the original and the Haar wavelet transformed insertion loss values as an example. As a matter of course, wavelet transformation results in discrete values, however, to make the corresponding points visible, they are graphically linked in Fig. 4.

The rule base using Daubechies wavelets did not give better results than the ones without any wavelet transformation, moreover, several additional errors occured. On the contrary, in case of Haar wavelets, accurate results arose for each of the 13 lines that produced valid results and one further line could be assessed, too, as it can be seen in the left hand side of Fig. 5.

In order to evaluate those lines that were previously not to be assessed, the new, Haar wavelets-type rule base was applied together with the stabilized KH rule interpolation. The 65 test lines were re-processed, thus the predictions became feasible in case of all lines. The predictions for the 13 wire pairs which were correctly evaluated previously remained valid, moreover, results of the predictions of 33 from the other 52 were correct, and 19 acceptable (in this contribution, results with a deviation of -1 from the correct values are considered as acceptable ones, all the others as incorrect) and there were no incorrect results.

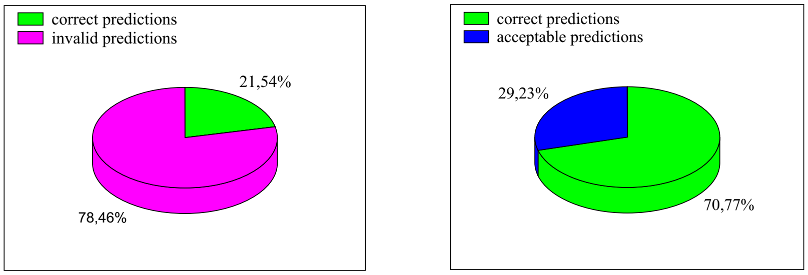


Fig. 5. Efficiency of the Haar wavelets based rule base alone (left) and supplemented with the stabilized KH rule interpolation (right).

The simplified "algorithm" of the construction of the predicting system is summarized as follows.

- Collection of insertion loss and bit rate data of wire pairs.
- Dividing the whole bit rate domain into groups (the more the number of the measured lines, the finer is the possible resolution) and clustering the measured values into these groups.
- Generation of several discrete values (6 in this case, however, other resolutions are examined by our ongoing investigations) from measured insertion loss functions by wavelet transformation (Haar wavelets are now recommended, though investigating other types of wavelets with other resolution levels are being in progress).
- Construction of fuzzy rule bases by clustered and wavelet transformed values.
- Wavelet transformation of the insertion loss function of the wire pair to be predicted.
- Prediction making by stabilized KH interpolation (can be made even if the input values can be found within the areas covered by antecedent fuzzy sets).

5 Conclusions

A novel performance prediction method based on interpolated fuzzy inference for telecommunications transmission lines and wavelet transformation of the values of the physical parameters influencing the performance was presented. The combination of the fuzzy rule interpolation and wavelet transformation was proposed in this paper in the first time. Wavelet transform was used for generating a coarse-grained view of the measured data, whereas the interpolation is applied for treating the sparseness of the rule bases. The method performed very well for the model system of the SHDSL connections, 52 predictions from 65 test cases were correct, and the other 19 were acceptable.

References

1. F. Lilik, J. Botzheim: Fuzzy based Prequalification Methods for EoSHDSL Technology, Acta Technica Jauriensis, Series Intelligentia Computatorica, Vol. 4. No. 1., pp.135-145, 2011.
2. F. Lilik, L. T. Kóczy: The Determination of the Bitrate on Twisted Pairs by Mamdani Inference Method, Issues and Challenges of Intelligent System and Computational Intelligence, Studies in Computational Intelligence, vol 530, 2014, pp.59-74, doi: 10.1007/978-3-319-03206-1_5
3. K. Balázs, L. T. Kóczy: Constructing Dense, Sparse and Hierarchical Fuzzy Systems by Applying Evolutionary Optimization Techniques, Applied and Computational Mathematics, Vol. 11, No. 1, 2012, pp. 81-101
4. F. Lilik, Sz. Nagy, L. T. Kóczy: Wavelet Based Fuzzy Rule Bases in Pre-qualification of Access Networks' Wire Pairs, IEEE Africon 2015, Addis Ababa, Ethiopia, 14-17 September 2015.
5. I. Daubechies: Ten Lectures on Wavelets, CBMS-NSF regional conference series in applied mathematics 61, SIAM, Philadelphia, 1992.
6. Ch. Christopoulos, A. Skodras, T. Ebrahimi: The JPEG2000 Still Image Coding System: An Overview, IEEE Trans. Consumer Electronics, Vol. 46, pp. 1103-1127 (2000). doi:10.1109/30.920468
7. Sz. Nagy, J. Pipek: On an economic prediction of the finer resolution level wavelet coefficients in electron structure calculations Phys. Chem. Chem. Phys., 2015, Accepted Manuscript doi:10.1039/C5CP01214G
8. A. Haar, Zur theorie der orthogonalen funktionen systeme, Math. Ann., Vol. 69, pp. 331-371 (1910).
9. L. T. Kóczy, K. Hirota: Approximate reasoning by linear rule interpolation and general approximation, International Journal of Approximate Reasoning, vol. 9, 1993, pp. 197-225, doi: 10.1016/0888-613X(93)90010-B
10. L. T. Kóczy, K. Hirota: Interpolative reasoning with insufficient evidence in sparse fuzzy rule bases, Information Sciences, vol. 71, 1993, pp. 169-201, doi: 10.1016/0020-0255(93)90070-3
11. D. Tikk, I. Joó, L. T. Kóczy, P. Várlaki, B. Moser and T. D. Gedeon: Stability of interpolative fuzzy KH-controllers, Fuzzy Sets and Systems, vol. 125, 2002, pp. 105-119, doi: 10.1016/S0165-0114(00)00104-4

State reduction methods for Fuzzy Cognitive Map to model regional waste management systems

Miklós F. Hatwágner[1], Adrienn Buruzs[2] and László T. Kóczy[3]

[1]Department of Information Technology, Széchenyi István University, Győr, Hungary
miklos.hatwagner@sze.hu
[2]Department of Environmental Engineering, Széchenyi István University, Győr, Hungary
buruzs@sze.hu
[3]Department of Automation, Széchenyi István University, Győr, Hungary and Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Hungary
koczy@sze.hu, koczy@tmit.bme.hu

Abstract. The authors have investigated the sustainability of Integrated Waste Management Systems (IWMS). These systems were modeled by Fuzzy Cognitive Maps (FCM), which are known as adequate fuzzy-neural network type models for multi-component systems with a stable state. The FCM model was designed of thirty-three factors to describe the real world processes of IWMS in as much detailed and as much accurately as possible. Although, this detailed model meets the requirements of accuracy, the presentation and explanation of such a complex model is difficult due to its size.

While there is a general consensus in the literature about a very much simplified model of IWMSs, detailed investigation lead to the assumption that a much more complex model with considerably more factors (components) would more adequately simulate the rather complex real life behavior of the IWMS.

As the starting point we used the thirty-three component model based on the consensus of a workshop of experts coming from all areas of the IWMS (operation, regulation, management, etc.) and the set goal was to find the most accurate real model that could be obtained by analyzing and properly reducing this – very likely too much detailed, or atomized – model.

In this paper, a new state reduction approach is presented. The practical aspects of the results gained by these methods are evaluated.

Keywords: fuzzy cognitive maps, integrated waste management system, state reduction methods.

1 Introduction

During the previous investigations [1] the method of FCM was applied to model regional waste management systems which are determined by six factors. As a validation of the simulation results [2] data were collected based on the relevant literature to set up a time series. This time series served as an input to the Bacterial Evolutionary Algorithm (BEA) which generated an optimal connection matrix producing the possibly

adfa, p. 1, 2011.

most similar time series to the original one obtained from the literature. Despite the expectations, the six-factor FCM model proved to be rather inaccurate in practice [3], and this is why a refined, more detailed model, containing thirty-three factors was developed [6] with the support of a group of experts.

After the thorough examination of both the basic (6 factors) and the detailed (33 components) models it became apparent that the two models were very different, in their respective complexities and concepts. For this reason, we assumed that an intermediate model containing less than thirty-three but more than six factors would be presumably able to describe the mechanism and action of a real IWMS with sufficient accuracy.

Table 1 introduces the main factors of the basic model and the thirty-three sub-factors of the detailed model.

Table 1. The identified sub-factors of the main factors and the concept IDs (CID) of them

Main factor	Sub-factor	CID
Technology (C1)	Engineering knowledge	C1.1
	Technological system and its coherence	C1.2
	Local geographical and infrastructural conditions	C1.3
	Technical requirements in the EU and national policy	C1.4
	Technical level of equipment	C1.5
Environment (C2)	Impact on environmental elements	C2.1
	Waste recovery	C2.2
	Geographical factor	C2.3
	Resource use	C2.4
	Wildlife (social acceptance)	C2.5
	Environmental feedback	C2.6
Economy (C3)	Composition and income level of the population	C3.1
	Changes in public service fees	C3.2
	Depreciation and resource development	C3.3
	Economic interest of operators	C3.4
	Financing	C3.5
	Structure of industry	C3.6
Society (C4)	Public opinion	C4.1
	Public health	C4.2
	Political and power factors	C4.3
	Education	C4.4
	Culture	C4.5
	Social environment	C4.6
	Employment	C4.7
Law (C5)	Monitoring and sanctioning	C5.1
	Internal and external legal coherence (domestic law)	C5.2
	General waste management regulation in the EU	C5.3
	Policy strategy and method of implementation	C5.4
Institution (C6)	Publicity, transparency (data management)	C6.1
	Elimination of duplicate authority	C6.2
	Fast and flexible administration	C6.3
	Cooperation among institutions	C6.4
	Improvement of professional standards	C6.5

On the basis of the detailed model, we might be able to support the strategic decision making process of the stakeholder in order to ensure the long-term sustainability of IWMS.

2 The investigated state reduction method

The idea of state reduction is similar to clustering but it can also be considered as a special, strongly generalized version of the state reduction technique of sequential circuits or finite state machines (see e.g. [10]). The methods construct clusters of factors and these clusters can be used later as factors of the reduced model. The members of clusters are selected based on their ‘similarity’. Two factors are considered similar, if their ‘distance’ is low. The distance of them have to be measured by an appropriate metric, and the applied metric differentiate the methods from each other. Different metrics and different distance values can also result in different clusters and reduced models. Several different metrics have been proposed e.g. in [7], but the basic idea of them is always the same and all versions use only the connection matrix and a threshold value of the maximum allowed distance. In this paper only one of the best solutions will be presented. Two factors C_i and C_j are considered similar, if the connections originating from them and leading to a third factor C_k, and also in the opposite direction have almost the same weights for all C_k, where $1 \leq i < j \leq n$, n is the number of factors, and $i \neq k, j \neq k, 1 \leq k \leq n$. At first, all clusters contain only one of the factors, but as soon as a similar factor is found, they will be merged. During the next steps, the similarity of all current cluster members must be measured to the next candidate factor.

The main properties of similarity are the following: 1) all factor is similar to itself (reflexivity), 2) if factor C_i is similar to C_j, then C_j is similar to C_i as well (symmetry). 3) But if C_i is similar to C_j and C_j is similar to C_k, then C_i is not always similar to C_k (non-transitive). It means that the state reduction method is a fuzzy tolerance relation [5, 8].

After the presentation of the basic idea, the precise description of the methods are given. First, the clusters are disjoint sets of factors, and each one of them contains only one factor. $K_i = \{C_i\}$ for every $i = 1 \dots n$ where K_i is the ith cluster, C_i is the ith factor (factors are often called ‘concepts’ in the FCM theory) and n is the number of factors in the model (thirty-three in the IWMS model). In the next steps all clusters will be appended by other factors, if possible. The ‘distance’ between the next cluster candidate and all current cluster members are measured by the chosen metric.

The presented metric calculates the normalized, squared Euclidean distance of the connections starting from factors C_i and C_j to C_k, where $i \neq j \neq k$, $i, j, k = 1 \dots n$. If this difference is below the threshold value (ε), the current factor is added to the cluster. The determined distance is normalized to [0, 1]. The applied metric is described more precisely by the following C-style pseudo-code (see **Fig. 1** and **Fig. 2**).

```
function isNear(i, j, eps, c)
  sum = 0;  // i, j = factor indexes, eps = ε
  for(k=0; k<n; k++) // n = number of factors
    if(k!=i and k!=j and !elementOf(k, c))
      dout = w(i, k)-w(j, k) // w(i, k) = $w_{ik}$
      sum = sum + dout * dout
      din = w(k, i)-w(k, j)
      sum = sum + din * din
  if(sum / ((n-2)*8) < eps)
    return true
  else
    return false
```

Fig. 1. Calculation of the distance of two concepts

```
function buildCluster(initialFactor, eps)
  c = {initialFactor}
  for(i=0; i<n; i++)
    if(i != initialFactor)
      member = true
      while(member and hasNextElement(c))
        j = nextElement(c)
        member = isNear(j, i, eps, c)
        if(member)
          c = c + {i}
  return c

function buildAllClusters(eps)
  clusters = {}
  for(i=0; i<n; i++)
    k = buildCluster(i, eps)
    if(!isElementOf(k, clusters))
      clusters = clusters + {k}
  return clusters
```

Fig. 2. Pseudo-code of the state reduction algorithm, Part 1

The state reduction is started by the `buildAllClusters` function (see **Fig. 2**). It requests the creation of each clusters by consecutive calling of the `buildCluster` function. The latter function sometimes produces the same clusters in different order, but `buildAllClusters` keeps only one of them. The distance of two factors are measured by `isNearA` or some of the other functions implementing different metrics.

When all the clusters are defined, the weights of the interconnections are defined by function `getWeight`. This function accepts two cluster arguments and provides the weight between these clusters. The return value is the average weight of connections among the factors of the specified clusters. The weight of self-loops are always zero according to the original FCM definition (see **Fig. 3**).

```
function getWeight(a, b)
  count = 0
  sum = 0
  while(hasNextElement(a))
    i = nextElement(a)
    while(hasNextElement(b))
      j = nextElement(b)
      if(i != j)
        count = count + 1
        sum = sum + w(i, j)
  if(count == 0)
    return 0
  else
```

```
    return sum/count
```

Fig. 3. Pseudo-code of the state reduction algorithm, Part 2

The value of ε must be in the [0, 1] interval and must be chosen appropriately in every single case, because it plays an important role in the reduction process. Too low values do not lead to models containing significantly fewer factors (clusters), thus they are not useful. But if the value of ε is too high, the model will be oversimplified and will not have the required accuracy. For example in an extreme case, when ε is 1, the whole model collapses and only one big sole cluster remains. The knowledge and experience of experts are needed to specify a meaningful ε value. In order to show the connection between ε and the number of factors (clusters) some interesting value pairs are collected in Table 2.

Table 2. The number of factors in the reduced connection matrix

ε	No. of factors
0.015	29
0.023	24
0.024	23
0.037	22
0.070	21
0.080	17

It must be emphasized here that all model reduction activities necessarily cause information loss, and the accuracy of simplified models are always lower. In the suggested method, there are three root causes of information loss:

1. The connections among concepts inside the same cluster are neglected. The representation of these connections would result in self loops which is not allowed according to Kosko's original idea.
2. Every causal relation needs one time step before their effect can be observed. Long pathes of concepts and interconnections may cause long delays. If more or less elements of these pathes become inside the same cluster, these delays partially disappear.
3. Since the model reduction method is based only on the connection matrix, the *getWeight* function cannot take into account the effect of possibly different (source) concept values on the connected (destination) concepts, because all such connections are represented by a single connection in the reduced model.

Despite all these possible problems, the proposed state reduction method performed well in several practical problems [9]. Furthermore, an exhaustive investigation is under fulfillment in order to analyze the behavior of the proposed method on statistical basis.

3 Results

In the next, the authors give an overview about and shortly analyze the driving forces and impact of IWMS upon the results of the state reduction method (Table 3, Table 4).

Table 3. An example of clusters as a result of state reduction (ε = 0.06)

Cluster ID	Reduced concepts
Q1	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.3 + C3.3 + C3.4 + C3.5 + C4.3 + C4.4 + C4.5 + C4.6 + C6.4
Q2	C1.1 + C1.3 + C2.1 + C2.3 + C2.5 + C4.2 + C4.3 + C4.4 + C4.5 + C4.7
Q3	C1.2 + C1.5 + C2.2 + C2.3 + C3.3 + C3.5 + C3.6 + C5.1
Q4	C2.4 + C2.5 + C2.6 + C4.4 + C4.5 + C4.6 + C5.1
Q5	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.3 + C2.5 + C3.3 + C3.4 + C3.5 + C4.3 + C4.5 + C6.4
Q6	C1.1 + C2.1 + C2.4 + C2.5 + C2.6 + C4.2 + C4.4 + C4.5 + C4.6 + C5.1
Q7	C1.1 + C1.2 + C1.3 + C1.4 + C2.5 + C3.1 + C3.3 + C3.4 + C3.5 + C4.3 + C4.4 + C4.5 + C4.6 + C6.1 + C6.2 + C6.3 + C6.4
Q8	C1.1 + C1.2 + C1.4 + C1.5 + C2.5 + C3.2 + C3.3 + C3.4 + C3.5 + C4.3 + C4.4 + C4.5 + C4.6 + C6.1 + C6.2 + C6.3 + C6.4
Q9	C1.1 + C1.2 + C1.4 + C1.5 + C2.3 + C3.3 + C3.4 + C3.5 + C3.6 + C4.3 + C5.3 + C6.4
Q10	C1.1 + C1.2 + C1.4 + C1.5 + C2.5 + C4.1 + C4.2 + C4.4 + C4.5 + C4.6 + C4.7 + C5.2 + C5.3 + C6.4
Q11	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.5 + C4.2 + C4.4 + C4.5 + C4.6 + C4.7 + C5.3 + C6.4
Q12	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.3 + C2.5 + C3.3 + C3.4 + C3.5 + C4.3 + C4.4 + C4.5 + C4.6 + C6.4
Q13	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.3 + C3.3 + C3.4 + C3.5 + C4.3 + C4.4 + C4.5 + C4.7 + C6.4
Q14	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.3 + C3.3 + C3.5 + C4.3 + C4.4 + C4.5 + C4.6 + C5.1
Q15	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.5 + C3.3 + C3.5 + C4.4 + C4.5 + C4.6 + C5.2 + C5.3 + C6.4
Q16	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.3 + C3.3 + C3.4 + C3.5 + C4.3 + C4.5 + C5.3 + C6.4
Q17	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.3 + C3.3 + C3.4 + C3.5 + C4.3 + C4.5 + C5.4 + C6.4
Q18	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.3 + C2.5 + C3.3 + C3.4 + C3.5 + C4.3 + C4.5 + C6.1 + C6.4
Q19	C1.1 + C1.2 + C1.3 + C1.4 + C2.3 + C3.3 + C3.4 + C3.5 + C4.3 + C6.1 + C6.2 + C6.3 + C6.4 + C6.5
Q20	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.3 + C3.3 + C3.4 + C3.5 + C4.3 + C4.4 + C4.5 + C6.3 + C6.4
Q21	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.3 + C3.3 + C3.4 + C3.5 + C4.3 + C4.4 + C4.5 + C6.4
Q22	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.3 + C3.3 + C3.4 + C3.5 + C4.3 + C4.4 + C4.5 + C6.3 + C6.4 + C6.5

Table 4. An example of clusters as a result of state reduction (ε = 0.08)

Cluster ID	Reduced concepts
Q1	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.1 + C2.2 + C2.3 + C2.4 + C2.5 + C2.6 + C3.2 + C3.3 + C3.4 + C3.5 + C3.6 + C4.1 + C4.2 + C4.3 + C4.4 + C4.5 + C4.6 + C4.7 + C5.2 + C5.3
Q2	C1.1 + C1.2 + C1.3 + C1.5 + C2.1 + C2.2 + C2.3 + C2.4 + C2.5 + C2.6 + C3.2 + C3.3 + C3.4 + C3.5 + C3.6 + C4.1 + C4.2 + C4.3 + C4.4 + C4.5 + C4.6 + C4.7 + C5.2 + C5.3
Q3	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.1 + C2.3 + C2.5 + C2.6 + C3.1 + C3.2 + C3.3 + C3.4 + C3.5 + C3.6 + C4.3 + C4.4 + C4.5 + C4.7 + C5.1 + C5.3 + C5.4
Q4	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.2 + C2.3 + C2.4 + C2.5 + C2.6 + C3.2 + C3.3 + C3.4 + C3.5 + C3.6 + C4.2 + C4.3 + C4.4 + C4.5 + C4.6 + C5.2 + C5.3 + C5.4 + C6.2 + C6.5
Q5	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.3 + C2.5 + C2.6 + C3.1 + C3.2 + C3.3 + C3.5 + C4.1 + C4.3 + C4.4 + C4.5 + C4.6 + C5.2 + C5.4

Q6	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.1 + C2.3 + C2.5 + C2.6 + C3.1 + C3.2 + C3.3 + C3.4 + C3.5 + C4.1 + C4.2 + C4.3 + C4.4 + C4.5 + C4.6 + C4.7 + C5.2 + C5.3 + C6.4
Q7	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.1 + C2.2 + C2.3 + C2.4 + C2.5 + C2.6 + C3.4 + C3.5 + C3.6 + C4.1 + C4.2 + C4.3 + C4.4 + C4.5 + C4.6 + C5.2 + C5.3 + C6.4 + C6.5
Q8	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.1 + C2.3 + C2.5 + C2.6 + C3.1 + C3.2 + C3.3 + C3.4 + C3.5 + C3.6 + C4.3 + C4.4 + C4.5 + C4.6 + C5.3 + C5.4
Q9	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.1 + C2.2 + C2.3 + C2.4 + C2.5 + C2.6 + C3.3 + C3.4 + C3.5 + C3.6 + C4.1 + C4.2 + C4.3 + C4.4 + C4.5 + C4.6 + C5.1 + C5.2 + C5.3
Q10	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.1 + C2.3 + C2.5 + C2.6 + C3.1 + C3.2 + C3.3 + C3.4 + C3.5 + C3.6 + C4.3 + C4.4 + C4.7 + C5.1 + C5.2 + C5.3 + C5.4 + C6.5
Q11	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.1 + C2.2 + C2.3 + C2.4 + C2.5 + C2.6 + C3.3 + C3.4 + C3.5 + C3.6 + C4.3 + C4.4 + C4.5 + C4.6 + C4.7 + C5.3 + C6.5
Q12	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.1 + C2.3 + C2.5 + C2.6 + C3.1 + C3.2 + C3.3 + C3.4 + C3.5 + C3.6 + C4.3 + C4.4 + C4.7 + C5.1 + C5.3 + C5.4
Q13	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.1 + C2.3 + C2.5 + C2.6 + C3.1 + C3.2 + C3.3 + C3.4 + C3.5 + C3.6 + C4.3 + C4.4 + C4.7 + C5.1 + C5.4 + C6.1
Q14	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.1 + C2.3 + C2.5 + C2.6 + C3.1 + C3.2 + C3.3 + C3.4 + C3.5 + C3.6 + C4.3 + C4.4 + C4.7 + C5.3 + C5.4 + C6.1 + C6.2 + C6.4
Q15	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.3 + C2.5 + C3.1 + C3.2 + C3.3 + C3.4 + C3.5 + C3.6 + C4.3 + C4.4 + C4.5 + C4.6 + C4.7 + C5.1 + C6.2 + C6.3
Q16	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.1 + C2.3 + C2.5 + C2.6 + C3.1 + C3.2 + C3.3 + C3.4 + C3.5 + C3.6 + C4.3 + C4.4 + C4.5 + C5.1 + C5.3 + C5.4 + C6.4
Q17	C1.1 + C1.2 + C1.3 + C1.4 + C1.5 + C2.1 + C2.2 + C2.3 + C2.5 + C2.6 + C3.2 + C3.3 + C3.4 + C3.5 + C3.6 + C4.3 + C4.4 + C4.5 + C5.1 + C6.5

As it can be seen from the above tables, there are several overlaps among the clusters. According to this, some of the factors are presented many times in the new models. The role of factors are described in Table 5 and Table 6.

Table 5. Appearance of factors in the clusters – the possible framework of a new IWMS (ε = 0.06)

CID	Role of factors	Nomination of factor	CID	Role of factors	Nomination of factor
C1.1	21	Engineering knowledge	**C1.4**	19	Technical requirements in the EU and national policy
C2.5	21	Wildlife (social acceptance)	**C3.3**	19	Culture
C4.3	21	Political and power factors	**C3.5**	19	Financing
C1.5	20	Technical level of equipment	**C1.2**	18	Technological system and its coherence
C1.3	19	Local geographical and infrastructural conditions	**C2.3**	18	Geographical factor

Table 6. Appearance of factors in the clusters – the possible framework of a new IWMS (ε = 0.08)

CID	Role of factors	Nomination of factor	CID	Role of factors	Nomination of factor
C1.1	**17**	Engineering knowledge	**C2.5**	**17**	Wildlife (social acceptance)
C1.2	**17**	Technological system and its coherence	**C3.5**	**17**	Financing
C1.3	**17**	Local geographical and infrastructural conditions	**C4.3**	**17**	Political and power factors

C1.5	**17**	Technical level of equipment	**C4.4**	**17**	Education
C2.3	**17**	Geographical factor	**C1.4**	**16**	Technical requirements in the EU and national policy

Integrated modeling requires not only the consideration of the technical and economic system elements, but also social, environmental, legal and institutional factors, furthermore their sub-factors. In these cases, to deal with the situations where data at hand are often insufficient for an entire quantitative analysis and the uncertainty is high, a series of non-quantifiable elements become important.

IWMSs are organized along spatial and temporal scales. While modeling the system, it leads to the appearance of the connections and interaction between its factors and sub-factors. The factors of these systems are connected via material, energy, money and information flows and form a complex phenomenon through legal regulation.

During studying the results of the state reduction methods, it could be recognized that the most commonly occurring factor in the new models (Table 5, Table 6) is the ‘Engineering knowledge’. Based on international experience it might be still surprising that the most important element in the system is ‘Engineering knowledge’. The factor has a determining role in the design and operation of the systems according to technical-economic-environmental considerations.

This combination of factors tells about the relationship of the sub-factors as parts of the systems and highlights the question ‘What is important in this system?’.

4 Conclusions

A new state reduction approach was introduced to make the otherwise too complex connection matrix of IWMS model easier to handle and understand. The factors of two simplified connection matrices was presented and evaluated. The presented reduction method will be further investigated in regard to its known shortcomings.

The state reduction approach proved to be good to combine different type of factors and create clusters. It thereby provides a comprehensive and more thoroughly understanding of an IWMS as a technical-economic-social-environmental system.

The conclusions based on the results of state reduction should be viewed together with existing scientific knowledge. In the next period, it is the authors’ intention to study further the assumptions, but also be open to insights gained from a systemic approach to deliver a method for decision making on sustainable regional waste management.

5 Future research

The authors’ purpose is to apply a modified version of the FCM. In the suggested method several connections that existed in the original model were neglected as their source and sink factors were represented by a single cluster of factors. The representation of these internal connections may lead to self-loops, which means non-zero elements in the main diagonal of the connection matrix. Usually this property of the matrix is not accepted [4], but real life systems query the justification of this theory. The usage of clusters also removes some delays of the original models, which can cause different simulation results or limit cycle behavior. The authors intention is to analyze the effect caused by self-loops and missing delays, then provide advanced model reduction techniques.

Acknowledgement. The authors would like to thank to TÁMOP-4.2.2.A-11/1/KONV-2012-0012, TÁMOP-4.1.1.C-12/1/KONV-2012-0017, to the Hungarian Scientific Research Fund (OTKA) K105529 and K108405 for the support of the research.

References

1. Buruzs, A., Pozna, R. C. and Kóczy, L. T. (2013). Developing Fuzzy Cognitive Maps for Modeling Regional Waste Management Systems. In Proceedings of the Third Int. Conference on Soft Computing Technology in Civil, Structural and Environmental Engineering, Paper 19, Tsompanakis, Y. (ed), Civil-Comp Press, Stirlingshire, UK.
2. Buruzs, A., Hatwágner, M. F., Pozna, R. C. and Kóczy, L. T. (2013). Advanced Learning of Fuzzy Cognitive Maps of Waste Management by Bacterial Algorithm. In Proceedings of IFSA World Congress and NAFIPS Annual Meeting, IEEE, pages 890-895.
3. Buruzs, A., Hatwágner, M. F. and Kóczy, L. T. (2014) Modeling integrated sustainable waste management systems by fuzzy cognitive maps and the system of systems concept, in: Technical Transactions series Automatic Control, accepted for publication, 2014
4. Carvalho, J. P. (2010) On the Semantics and the Use of Fuzzy Cognitive Maps in Social Sciences, WCCI 2010 IEEE World Congress on Computational Intelligence, July, 18-23, 2010 - CCIB, Barcelona, Spain
5. Das, M., Chakraborty, M. K., Ghoshal, T. K. (1998) Fuzzy tolerance relation, fuzzy tolerance space and basis, in: Fuzzy Sets and Systems, Vol. 97, Issue 3, pages 361-369.
6. Hatwágner, M. F., Buruzs, A., Földesi, P. and Kóczy, L. T. (2014) Strategic decision support in waste management systems by state reduction in FCM models. ICONIP2014 in prep.
7. Hatwágner, M. F., Kóczy, L. T. (2015) Parameterization and concept optimization of FCM models, submitted to FUZZ-IEEE 2015.
8. Klir, G. J., Folger, T. A. (1987) Fuzzy Sets, Uncertainty and Information, Prentice Hall.
9. Papageorgiou, E.I., Hatwágner, M.F., Buruzs, A., Kóczy, L.T., A new clustering approach for designing reduced Fuzzy Cognitive Map models for decision making: application on Hungarian Waste Management System, submitted to Journal of Applied Soft Computing
10. Paull, M. C. (1959) Minimizing the Number of States in Incompletely Specified Sequential Switching Functions, in: IRE Transactions on Electronic Computers, Vol. EC-8, Issue 3, pages 356 – 367.

On the Use of F-transform on the Reduction of Concept Lattices

Petra Hodáková and Nicolás Madrid

University of Ostrava, Centre of Excellence IT4Innovations,
Institute for Research and Applications of Fuzzy Modeling,
30. dubna 22, 701 03 Ostrava 1, Czech Republic
{nicolas.madrid,petra.hodakova}@osu.cz

Abstract. In this paper, we show that F-transform can be used to reduce relational databases. Subsequently, we show that the respective concept lattice is reduced significantly as well. Moreover, we present a clarifying example of the procedure.

Keywords: F-transform, Fuzzy Sets, Fuzzy Concept Analysis, Knowledge Reduction.

1 Introduction

Fuzzy Formal Concept Analysis deals with the processing of imprecise knowledge in information systems [2, 3]. In this theory, the information of a relational databases is represented in terms of a complete lattice where its elements are called concepts. However, despite the information represented by the concept lattice is valuable, the complexity and the size (which increases exponentially with respect to the size of the relational database) makes the use of this theory impractical in many applied tasks. For this reason, recent approaches have dealt with Knowledge Reduction in relational databases to simplify the formal concept analysis of them [1].

On the other hand, F-transforms [4] is a theoretical tool that has shown its effectiveness on representing the information of signals (like temporal series, images, etc.) to a vector of few components. This paper applies F-transforms (based on residauted lattice) to Knowledge Reduction. Specifically, we begin by showing that objects (or attributes) in a relational database can be grouped in a new set of objects (or attributes). Then, we transfer the information of the original database to another where objects are given by the grouping previously mentioned. The transfer of information is given by F-transforms and therefore, there are two possible new relational databases.

This paper has the following structure. In Section 2 we recall briefly the theories of fuzzy property-oriented concept lattices and F-transforms. Then, in Section 3 we describe the reduction of relational tables by means of F-Transforms. Moreover, we illustrate the consequences of the reduction in concept lattices with an example. Finally, in Section 4 we present conclusions and future work.

2 Preliminaries

2.1 F-transforms on residuated lattices

In this section, we briefly recall the basic definitions and the main principles of F-transforms based on operations of a residuated lattice [4]. Let $(L, \leq, \&, \rightarrow)$ be a residuated lattice. A fuzzy partition of a finite set $\mathcal{U}$ is a set of L-fuzzy sets on $\mathcal{U}$ $\mathcal{A}_1, \ldots, \mathcal{A}_n$ fulfilling the covering property namely, for all $x \in \mathcal{U}$ there exists $k \in \{1, \ldots, n\}$ such that $\mathcal{A}_k(x) > 0$. The membership functions $\mathcal{A}_k(x)$, $k = 1, \ldots, n$ are called the *basic functions.*

Definition 1. *Let $f : \mathcal{U} \to L$ be a function and $\mathcal{A}_1, \ldots, \mathcal{A}_n$, with $n \leq |\mathcal{U}|$, be basic functions which form a fuzzy partition of $\mathcal{U}$. We say that the n-tuple of real numbers $\mathbf{F}^{\uparrow}_n[f] = [F^{\uparrow}_1, \ldots, F^{\uparrow}_n]$ is the (direct) $F^{\uparrow}$-transform of f w.r.t. $\mathcal{A}_1, \ldots, \mathcal{A}_n$ if*

$$F^{\uparrow}_k = \bigvee_{x \in \mathcal{U}} (\mathcal{A}_k(x) \,\&\, f(x)). \tag{1}$$

Moreover, we say that the n-tuple of real numbers $\mathbf{F}^{\downarrow}_n[f] = [F^{\downarrow}_1, \ldots, F^{\downarrow}_n]$ is the (direct) $F^{\downarrow}$-transform of f w.r.t. $\mathcal{A}_1, \ldots, \mathcal{A}_n$ if

$$F^{\downarrow}_k = \bigwedge_{x \in \mathcal{U}} (\mathcal{A}_k(x) \to f(x)). \tag{2}$$

The elements $F^{\uparrow}_1, \ldots, F^{\uparrow}_n$ and $F^{\downarrow}_1, \ldots, F^{\downarrow}_n$ are called *components* of the $F^{\uparrow}$-transform and $F^{\downarrow}$-transform, respectively.

The following lemma ([4]) shows that the components of the $F^{\uparrow}$-transform ($F^{\downarrow}$-transform) are lower mean values (upper mean values) of an original function which give least (greatest) element to certain sets.

Lemma 1. *Let $f : \mathcal{U} \to L$ be a function and $\mathcal{A}_1, \ldots, \mathcal{A}_n$, with $n \leq |\mathcal{U}|$, be basic functions which form a fuzzy partition of $\mathcal{U}$. Then the k-th component of the $F^{\uparrow}$-transform is the least element of the set*

$$S_k = \{a \in L |\ \mathcal{A}_k(x) \leq (f(x) \to a) \textit{ for all } x \in \mathcal{U}\}$$

and the k-th component of the $F^{\downarrow}$-transform is the greatest element of the set

$$T_k = \{a \in L |\ \mathcal{A}_k(x) \leq (a \to f(x)) \textit{ for all } x \in \mathcal{U}\}$$

where $k = 1, \ldots, n$.

2.2 Fuzzy property-oriented concept lattices

In this section we recall briefly a simplification of property-oriented concept lattices introduced in [2, 3]. So, because of the lack of space, here we restrict to residuated lattices instead of adjoin triples. The notion of fuzzy property-oriented context is defined below.

Definition 2. *Let $(L, \leq, \&, \to)$ be a residuated lattice. A* context *is a tuple (A, B, R) such that A and B are non-empty sets (usually interpreted as attributes and objects, respectively), R is an L-fuzzy relation $R\colon A \times B \to L$.*

From now on, we fix a context (A, B, R). The mappings ${}^{\uparrow_\Pi} \colon L^B \to L^A$ and ${}^{\downarrow^N} \colon L^A \to L^B$ are defined, for $g \in L^B$ and $f \in L^A$ as, $g^{\uparrow_\Pi}$ and $f^{\downarrow^N}$, where

$$g^{\uparrow_\Pi}(a) = \bigvee_{b \in B} R(a,b) \,\&\, g(b)$$

$$f^{\downarrow^N}(b) = \bigwedge_{a \in A} R(a,b) \to f(a)$$

It is not difficult to prove that $({}^{\uparrow_\Pi}, {}^{\downarrow^N})$ forms an isotone Galois connection (also known as adjunction) and, therefore, ${}^{\uparrow_\Pi \downarrow^N} \colon L^B \to L^B$ is a closure operator and ${}^{\downarrow^N \uparrow_\Pi} \colon L^A \to L^A$ is an interior operator. A concept is a pair of mappings $\langle g, f \rangle$, with $g \in L^B, f \in L^A$, such that $g^{\uparrow_\Pi} = f$ and $f^{\downarrow^N} = g$, which will be called *fuzzy property-oriented concept*. In that case, g is called the *extent* and f, the *intent* of the concept. The set of all these concepts will be denoted as $\mathcal{F}_{\Pi N}$.

Definition 3. *The associated* fuzzy property-oriented concept lattice *to the context (A, B, R) is defined as the set*

$$\mathcal{F}_{\Pi N} = \{\langle g, f \rangle \in L^B \times L^A \mid g^{\uparrow_\Pi} = f \text{ and } f^{\downarrow^N} = g\}$$

in which the ordering is defined by $\langle g_1, f_1 \rangle \preceq \langle g_2, f_2 \rangle$ iff $g_1 \preceq_2 g_2$ (or equivalently $f_1 \preceq_1 f_2$).

3 Reducing the size of Relational Tables.

Throughout this section we consider a frame (A, B, R) and a residuated lattice $(L, \leq, \&, \to)$. The idea underlying in the reduct is the creation of two smaller relational tables $R^{\uparrow}$ and $R^{\downarrow}$ that keep as much information from R as possible. In order to reduce the size of the table is needed to reduce either the number of attributes or the number of objects. In this paper we focus on objects. In this way, we define a new set of objects $\overline{B}$ that can be considered as a set of fuzzy sets that group objects according to certain attributes in A. For instance, consider a relational table where objects are people and attributes are physical features of them. Then, we could group people according to their high and then, to define the following set of "new" objects $\{B_1 = VerySmall, B_2 = QuiteSmall, B_3 = Medium, B_4 = QuiteTall, B_5 = VeryTall\}$. To conclude the reduction, we only need to define the relations $R^{\uparrow}$ and $R^{\downarrow}$ between the new set of objects and the original set of attributes. For such a task we consider direct F-transforms. In this framework, each basic function from the chosen fuzzy partition determines a new object and the value assigned to it by the direct F-transform determines the value of the relation.

To define the basic functions (and then also the set of new objects) let us consider firstly, a fuzzy partition of L given by fuzzy sets $\{\mathcal{L}_k \colon k \in \{1, \dots, n\}\}$ and secondly, a subset of attributes $\overline{A} \subseteq A$. Then the fuzzy partition $\overline{B} = \{B_{k\overline{a}} \colon k \in \{1, \dots, n\}$ *and* $\overline{a} \in \overline{A}\}$ of B is defined by

$$B_{k\overline{a}}(b) = \mathcal{L}_k(R(b, \overline{a})), \quad b \in B. \tag{3}$$

Note that the fuzzy partition $\overline{B}$ groups original objects in fuzzy sets according to their relation with attributes in $\overline{A}$. Moreover, note the number of basic functions (i.e., the number of new objects) is $k \cdot |\overline{A}|$. So the size of the new set of objects depends on the number of attributes considered to define the partition. Once the fuzzy partition is fixed, we can define the following two L-fuzzy relational tables $R^{\uparrow}$ and $R^{\downarrow}$ between $\overline{B} = \{B_{k,\overline{a}} \colon k \in \{1, \dots, n\}$ *and* $\overline{a} \in \overline{A}\}$ and A as follows:

$$\begin{aligned}
R^{\uparrow} \colon & \overline{B} \times A \to L \\
& (B_{k,\overline{a}}, a) \mapsto \bigvee_{b \in B} B_{k\overline{a}}(b) \,\&\, R(a, b) \\
R^{\downarrow} \colon & \overline{B} \times A \to L \\
& (B_{k,\overline{a}}, a) \mapsto \bigwedge_{b \in B} B_{k\overline{a}}(b) \to R(a, b)
\end{aligned} \tag{4}$$

Note that original objects are used to define the values of the new ones. Finally, the reduction of the concept lattice given by the original frame (A, B, R) is the pair of concept lattices associated to the frames $(A, \overline{B}, R^{\uparrow})$ and $(A, \overline{B}, R^{\downarrow})$. Below we show how the procedure works in a simple example.

	$HighPower$	$BigSpace$	$HighConsume$	$Expensive$	$Sport$	$Familiar$
b_1	1	0.2	1	0.8	1	0
b_2	1	1	0.8	1	0.6	1
b_3	0.6	0.8	0.4	0.6	0.2	0.6
b_4	0.8	0.6	0.6	0.6	0.6	0.6
b_5	0.6	0.4	0.2	0.6	0.2	0.2
b_6	0	0.2	0	0.2	0	0
b_7	0.8	0.2	0.8	0.8	0.8	0
b_8	1	1	1	1	0	1
b_9	0.6	1	0.4	0.6	0	1
b_{10}	0.6	1	0.6	0.6	0	1
b_{11}	0.6	0.6	0.4	0.4	0	0.6
b_{12}	0.2	0.4	0.4	0.2	0	0.2
b_{13}	0.8	0	0.8	1	0.8	0

Fig. 1. A car relational database.

Example 1. Let us consider the L-relational table in Figure 1 that relates types of cars (objects) with features (attributes). For the sake of simplicity, let L be the unit interval $[0, 1]$ represented by finite set $L = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$, and the adjoint pair considered for the reduction and the construction of the concept lattice is the one given by the Gödel connectives. Let us consider the partition $\{\mathcal{L}_1, \mathcal{L}_2\}$ of L given by:

x	0	0.2	0.4	0.6	0.8	1
$\mathcal{L}_1(x)$	0	0.2	0.4	0.6	0.8	1

x	0	0.2	0.4	0.6	0.8	1
$\mathcal{L}_2(x)$	1	0.8	0.6	0.4	0.2	0

Now, for the sake of simplicity let us consider just the attribute $Familiar \in A$ to make the reduct. Then, from Equation (3), we have that the partition of the set of objects B with respect to the attribute $Familiar$ and the partition $\{\mathcal{L}_1, \mathcal{L}_2\}$ of L is given by the following two fuzzy sets

b	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
$B_{1\overline{a}}(b)$	0	1	0.6	0.6	0.2	0	0	1	1	1	0.6	0.2	0

b	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
$B_{2\overline{a}}(b)$	1	0	0.4	0.4	0.8	1	1	0	0	0	0.4	0.8	1

Note that partitions $B_{1\overline{a}}$ and $B_{2\overline{a}}$ above represent the fuzzy sets of cars that are familiar and non familiar, respectively. Thus, it has sense that the two new objects in the new tables are denoted by $FamCars$ and $NonFamCars$. The new relational tables are given by F-transforms (4) as follows

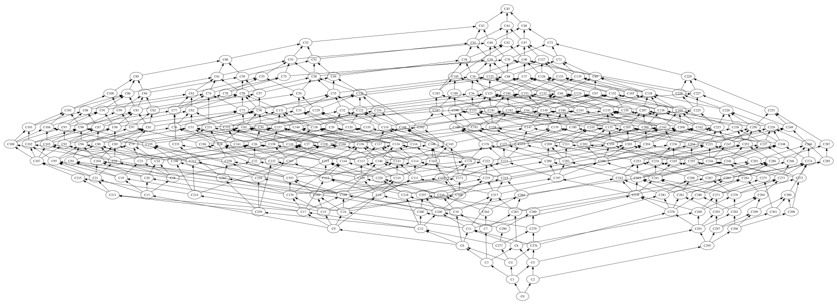
$R^{\uparrow}$	$HighPower$	$BigSpace$	$HighConsume$	$Expensive$	$Sport$
$FamCars$	1	1	0.8	1	0.6
$NonFamCars$	1	0.4	1	1	1

$R^{\downarrow}$	$HighPower$	$BigSpace$	$HighConsume$	$Expensive$	$Sport$
$FamCars$	0.6	1	0.4	0.6	0
$NonFamCars$	0	0	0	0	0

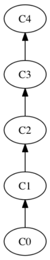
The tables above can be interpreted as follows. Tables $R^{\uparrow}$ and $R^{\downarrow}$ represent the possibility and necessity, respectively, of a familiar car (in some degree) to have a certain attribute. So $R^{\uparrow}(FamCars, a)$ and $R^{\downarrow}(FamCars, a)$ represent an upper and a lower bound, respectively, of the value $R(b, a)$ for any familiar car $b \in B$, i.e., $R^{\uparrow}(FamCars, a) \geq B_{1\overline{a}}(b) \,\&\, R(b, a)$ and $R^{\downarrow}(FamCar, a) \leq B_{1\overline{a}}(b) \to R(b, a)$.

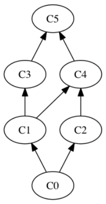
It is interesting to mention that from the interpretability above, we can infer from the tables $R^{\uparrow}$ and $R^{\downarrow}$ that a familiar car must have a big space because $R^{\uparrow}(FamCars, BigSpace) = R^{\downarrow}(FamCars, BigSpace) = 1$. Moreover, the familiar cars are quite powerful and expensive as well as $R^{\downarrow}(FamCars, HighPower) = R^{\downarrow}(FamCars, Expensive) = 0.6$.

The concept lattice of the original relational table of Example 1 has 302 concepts and it is given by the following Hasse diagram



However, the concept lattices of the tables reduced by our procedure have only 5 and 6 concepts, respectively.





4 Conclusion and Future Works

In this paper we have presented the reduction of relational tables aimed to keep as much information from the original table as possible. Our future work is to apply the reduct based on the ordinary F-transforms and measure, determine and/or bound the information which is on one hand lost by the reduction and on the other hand kept by the reduction.

Acknowledgments. The research was supported by the European Regional Development Fund by projects (CZ.1.05/1.1.00/02.0070) and (TIN12-39353-C04-04).

Aditionally, we want to thank Eloisa Ramírez-Poussa for her help.

References

1. M E. Cornejo, J. Medina, E. Ramírez-Poussa. Attribute reduction in multi-adjoint concept lattices. *Information Sciences*, 294: 41-56, 2015.
2. J. Medina. Towards multi-adjoint property-oriented concept lattices. *Lecture Notes in Artificial Intelligence*, 6401:159–166, 2010.
3. J. Medina. Multi-adjoint property-oriented and object-oriented concept lattices. *Information Sciences*, 190:95–106, 2012.
4. I. Perfilieva. Fuzzy transforms: Theory and applications. *Fuzzy Sets and Systems*, 157: 993–1023, 2006.

Possible Applications of Fuzzy Relational Calculus for Some General Problems of Recommender Systems

Alex Tormási[1], Brigitta Szi[1], Péter Földesi[2], Dávid Zibriczky[3] and László T. Kóczy[1,4]

[1]Department of Information Technology, Széchenyi István University, Győr, Hungary
{szi.brigitta,tormasi,koczy}@sze.hu
[2]Department of Logistics and Forwarding, Széchenyi István University, Győr, Hungary
foldesi@sze.hu
[3]ImpressTV, Budapest, Hungary
david.zibriczky@impresstv.com
[4]Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary
koczy@tmit.bme.hu

Abstract. The main contribution of this paper is to overview and discusses possible applications of fuzzy relational calculus to solve some issues and challenges of recommender systems. The presented ideas are targeting the most essential aspects of these problems, the knowledge representation and handling.

Keywords. fuzzy relations; recommender systems; implicit and explicit feedback; cold start problem; hybrid filtering

1 Introduction

Recommender systems are none other, than information filtering algorithms, that help users in the discovery of items in the multitude of choices. Personalized recommendations reduce the time the user spent for looking for relevant items and increase the likelihood of meeting the user's expectations. Recommender systems could be considered as cognitive info communication systems [1], which decrease the cognitive load and increases the mathability [2] of the users, extends the users' ability to filter out and access relevant content.

By the assistance of content discovery the user satisfaction may increase, consequently recommender systems have also impact from the business point of view. Because of this, recommender systems became more and more popular among both the businesses and end-users in the last decade. Recommender systems shouldn't consider the maximization of key performance indicators for the businesses only, but finding the trade off between accuracy, coverage, diversity and serendipity.

This paper is organized as follows: after the Introduction in Section 2 some of the main difficulties of recommender systems are outlined. In Section 3 there is a short overview of fuzzy relational calculus with possible applications in recommender systems. Fuzzy methods in recommender systems are briefly summarized in Section

4. This is followed by some ideas and aspects of fuzzy relational calculus in recommender systems. Summary of the paper and the main phases of future work are outlined in Section 6.

2 Issues and Challenges of Recommender Systems

There are several influencing factors that make difficult to fulfil all of the requirements for an effective recommender system. One of the key challenges is the proper interpretation of user activities for user profiling. In practice two types of user interaction is distinguished. We consider an interaction as "explicit feedback", when the user expresses his preference over an item intentionally (e.g. he gives a rating 4 for a movie). Explicit feedbacks have significant information about the preference since it can be quantified in the algorithms. The typical examples of explicit feedbacks are ratings, likes, dislikes and adding contents to favourites. The other type of interactions is the "implicit feedback" that refers to all kind of interactions that cannot be interpreted precisely in terms of preference. For example, a profile view interaction is considered as implicit feedback because it has no explicit meaning about how much the user likes or dislikes it. The typical examples of implicit feedbacks are viewing a profile, buying a product or watching a movie. Based on another approach, the difference between implicit and explicit feedback is that implicit feedbacks are generated before consuming the items while explicit feedbacks are made after the interaction with the item. In practice the collection of explicit data is more difficult, because it requires the intention and some efforts from the user to express its preference over the items. By contrast of that implicit feedbacks are much easier, because it is just a tracking of user browsing on the site. The consequence of these properties is that on one hand explicit data is more meaningful but it is less provided, on the other hand implicit data is less meaningful but it is provided with a higher level of magnitude.

Main properties of implicit and explicit feedback according to [3] and [4] are summarized as follows: It could be easy to collect implicit feedback from user interactions, but it is noisy, difficult to interpret and has a low accuracy. Explicit feedback has lower availability, possibly as a result of the increased cognitive information processing it requires, but has no such a reward, which would directly motivate the users to be involved. It is possible to determine both positive and negative preferences from explicit feedback, but it is dependent from the context [3] and could be noisy [5-7]. Since explicit and implicit feedbacks are representing the same consumer preferences, there must be a relation between them; D. Parra et al. used logistic regression in order to extract explicit feedback from implicit feedback [8].

Another practical problem is that the preference of users is changing over time, the interpretation of an interaction strongly depends on the context (e.g. the time of the day or the device that the user is using). The deeper understanding of the problem depends on the knowledge extracted from both the behaviour and cognitive processes of the users and the processes from industrial or commercial point of view including the advantages and disadvantages of applied technologies. In order to recommend items for the users, their profiles should be known. Usually it is possible through collecting and processing implicit and explicit feedback, which could be considered as different projections of the same user preferences.

Another challenge is solving the cold-start problem. A recommendation problem is considered as cold-start problem if neither explicit feedbacks nor implicit feedbacks are provided for an item or user. There are two kinds of cold-start problems [9-11], first is when a recommendation should be generated for a new user (with limited or without any previous knowledge of his preferences or patterns) and the other is when a new item appears in the system. To overcome these problems, content-based filtering (CBF) methods were introduced [9-12]. Metadata is essential to enhance the user and item models for better recommendation. At this stage basic data (such as gender, age group) are available for the system, but those could be ambiguous and it does not conclude that the user will have the typical preferences built from the data bank. A similar case is when a new item gets available and only basic information is ready such as categories, product descriptions or tags, which could be still misclassified. Content-based filtering techniques are targeting to use meta data to create more acceptable recommendation, but if the mate data is not well structured, the system will not perform well. The missing metadata might also lead to problems; when an item does not have proper description and/or tags some systems consider it not as missing data, but connects it with a negative property, which means that the missing tags are considered as features that do not apply to the item.

As the user interacts with the system, implicit or explicit feedbacks are collected about his behaviour. Analogously to new items, the users start to consume it and generating feedbacks for that. The increasing number of interactions improves the accuracy of user or item models. However it is generally true that more feedbacks results better models, there is a theoretical saturation point where an additional increment of the number of feedbacks doesn't result significant improvement in accuracy. The period between cold-start and saturation point is called "warm-up" period. It depends on the algorithm whether the user is in "warm-up" period, the key challenge is to reduce the amount of data required by the recommender algorithm.

The user interactions are not only used for personalization, but the extraction of user behavioural patterns, that called collaborative filtering (CF) [13]. Collaborative filtering methods recommend items based on what the similar users consumed. An advantage of collaborative filtering against the content-based filtering is that it is capable to extract behavioural patterns that cannot be explained by metadata. Conventional collaborative filtering can be powerful when a clear separation of user preferences is observed in consumption patterns. Usually it is not the case because the user preferences are usually mixed or fuzzy. The disadvantage of collaborative filtering methods is that they are not capable to solve the cold-start problem and performs.

To combine the advantages of collaborative- and content-based filtering, hybrid filtering was introduced [14]. Hybrid filtering methods are more complex than single collaborative- or content-based filtering methods, but offers better accuracy by solving cold-start problem and extracting consumption patterns at a time. Another advantage of hybrid filtering is that misclassified metadata can be easier detected (e.g. a movie labelled "action" mainly consumed by "romantic" movie fans), furthermore similarity between different tags can be evaluated. However the conventional hybrid filtering has many advantages, it is still difficult to handle missing information and less meaningful implicit data sets.

The properties of computational intelligence techniques (like fuzzy methods, meta heuristics) enable them to properly handle some of the described problems. There are number of works related to the application fuzzy methods in recommender systems and the aim of this work is to give a brief overview of these and look for some possible new perspectives.

3 Fuzzy Relations and Basic Operations

Similarly to crisp and fuzzy sets [15], the fuzzy relations could be interpreted as generalized form of crisp relations, where the connection between items of two or more (discrete or continuous) sets could be expressed by a membership degree [16–17]. If we consider relation *R* between sets X_1, X_2, …, X_n, then the formal description of the fuzzy relations is as follows:

$$R(X_1, \ldots, X_n) \subseteq X_1 \times \cdots \times X_n \, , \tag{1}$$

$$R(x_1, \ldots, x_n) = \mu R\langle x_1, \ldots, x_n \rangle . \tag{2}$$

Let *R* be a fuzzy relation over the X_1, …, X_n, then $[R \downarrow Y]$ is the projection of the relation on the *Y* multi sets formally:

$$[R \downarrow Y](\underline{y}) = \max_{\underline{y} \prec \underline{x}} R(\underline{x}) \cdot \tag{3}$$

The cylindrical extension could be considered as some sort of inverse operation of the above defined projection operator. It is marked as $[R \uparrow X - Y]$, where *R* is a fuzzy relation and *X*, *Y* are multi sets and the values are calculated for each *x*, where $\underline{y} \prec \underline{x}$:

$$[R \uparrow X - Y](\underline{x}) = R(\underline{y}) \cdot \tag{4}$$

The cylindrical closure (5) is similar to cylindrical extension, but it uses the intersection of multiple projections.

$$\mathrm{cyl}\{P_i\}(\underline{x}) = \min_{i \in I} [p_i \uparrow X - Y_i](\underline{x}) , \tag{5}$$

where P_i is a projection defined over Y_i multi set. $\{P_i | i \in I\}$ is a set of projections of *R* fuzzy relation defined over set *X*.

There is possibility for that nor do can the cylindrical extension and closure restore the original fuzzy relation, which situation can be illustrated as in Fig. 1.

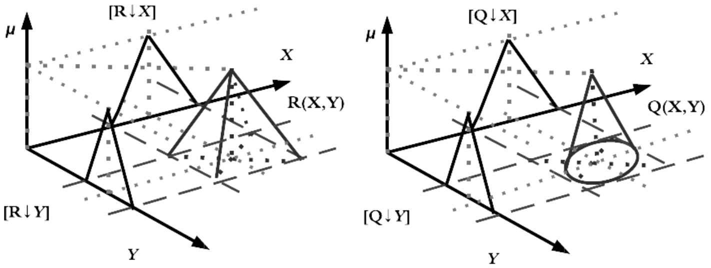


Fig. 1. Example of information loss during projection by two distinct fuzzy relations

The max-min composition of *P*(*X*, *Y*) and *Q*(*Y*, *Z*) binary fuzzy relations returns with *R*(*X*, *Z*) binary fuzzy relation, which is associative, its inverse is identical with the inverse relations' reverse composition, but does not satisfy the conditions of commutativity. It is easy to see it can be generalized to any t-norm and t-conorm (or s-norms) pairs. Formally described:

$$R(x,z)=[P\circ Q](x,z)=\max_{y\in Y}\min[P(x,y),Q(y,z)]\cdot \tag{6}$$

The relational connection operator is similar to the max-min composition; it produces an *R*(*X*, *Y*, *Z*) ternary fuzzy relation from the (relational) connection of *P*(*X*, *Y*) and *Q*(*Y*, *Z*) binary fuzzy relations and it also could be generalized to any t-norm and t-conorm pairs, formally:

$$R(x,y,z)=[P*Q](x,y,z)=\min[P(x,y),Q(y,z)]. \tag{7}$$

The fuzzy similarity measure of vectorvalued fuzzy (VVF) sets proposed and detailed in [18], [19] and [20] could be also applied for fuzzy relations. The similarity relation *S*(*R*, *Q*) of *R* and *Q* relations over *X* multi set is:

$$S_{RQ}=M_{R\equiv Q}, \tag{8}$$

where

$$R\equiv Q=(R\cap Q)\cup(\overline{R}\cap\overline{Q})\cdot \tag{9}$$

4 Fuzzy Methods in Recommender Systems

In the last couple of years, several solutions were proposed for the application of fuzzy methods for recommendation problems. In the evaluation of recommender system methods hybrid filtering seemed to be the most effective approach to overcome cold-start problem and exploit behavioural patterns that couldn't be explained solely metadata. Cornelis et al. applied fuzzy relations first in user and item similarities to improve the accuracy of conventional hybrid filtering methods [21].

Later, additional various fuzzy neighbour methods with the combination of collaborative and content-based filtering were introduced in [22-25].

One of the key challenges of recommender systems is to overcome the lack of information for user profiling. To address uncertainty due to vagueness Perez worked out a method by applying fuzzy preference relation that is capable to provide better recommendations for users with a few events [26], Zenebe proposed a general framework for discovering, interpreting and visualizing user preferences with fuzzy set theories [27]. The improvement of user warm-up period were also studied by Porcel and Herrera-Viedma, in [28] they presented a fuzzy linguistic recommendation strategy to improve the acquisition of user profile. Nilashi discussed the usability of fuzzy techniques in multi-criteria recommendation problem to provide better profile models [29].

In order to reduce the uncertainty of preference modelling fuzzy theories were also used for clustering methods. Nadi proposed a fuzzy clustering technique that captures user's behaviours on websites and provides more dynamic recommendations [30]. Liu and Gao studied the interpretation of user intentions with low amount of user actions, they proposed a recommendation solution by the application of fuzzy cluster analysis and cognitive maps [31]. Birtolo and Ronca published a study about two clustering collaborative filtering algorithms with the application of fuzzy logic. They measured a significant improvement in coverage of recommendations while the accuracy remained the same [32].

Several fuzzy-based recommendation methods were addressed to practical problems. Lu proposed a framework that helps students to find learning materials. For that he applied a multi-attribute evaluation method to capture the students' preferences and a fuzzy matching method to find the most suitable materials [33]. For telecommunication domain Wu designed a solution that deals with tree-based structure of contents by using fuzzy similarity measure [34]. Castro-Schez introduced a prototype of recommender system for B2C e-commerce businesses. They proposed a method that capable to deal with vague search preferences and provide fuzzy rule-based personalized recommendations of products [35]. Another application of fuzzy logic in e-commerce was proposed by Ramkumar, who introduced an automatic scoring for the reviews on products for spam detection [36]. Cornelis et al. addressed a method for the modelling of "one-and-only" items (the items that cannot be repetitive sold, e.g. houses). They applied fuzzy logic to extend existing collaborative filtering method and overcome the lack of collaboration [37]. García-Crespo applied fuzzy logic to provide personalized portfolio recommendations considering both financial attributes of investments and psychological aspects [38]. For the recommendation of candidates of political elections, Terán introduced a fuzzy clustering based method [39] and Dyczkowski proposed a voter preference modelling by intuitionistic fuzzy sets [40].

5 Possible Applications of Fuzzy Relations in Recommender Systems

The most trivial situation when a news site tries to categorize its viewers: some users tend to read latest news frequently no matter what kind of topic does it have, while older articles are read only according to his more specific preferences, which is suppressed by the data generated during reading the latest news. The main issues with profiling users and items were described in Section 2. In some situations there is no obvious way to define distinct clusters. The properties of fuzzy relations may help to overcome these problems. The fuzzy tolerance and equivalence relations have different mathematical properties, which makes it reasonable to investigate them in the context of recommendation, more specifically user and item preferences. An expression of the users/items and the clusters with fuzzy mathematics may have good results, since it is possible for a user (or item) to be part of various overlapping clusters. All the collected and non-collected (recommendation process related) data from the user are expressing his/her opinion and preferences over item or items. Both implicit, explicit feedback and meta data could be considered as sets of projections of a high-dimensional fuzzy relation; From this point of view it is obvious to assume that there hidden preferences of the user, which are not expressed in the collected data, but it does not require it for further processes. The composition and connection operators (or their generalized forms) defined over binary fuzzy relations (described in Section 3) could be also used to determine or estimate connections between various spaces of feedbacks and meta data.

Another possible application of fuzzy relations in user profiles are the following: consider a (global) fuzzy relation describing the users' basic data (e.g. gender, age group multi set) and their item consumption data (e.g. views, tags multi set). Assume a new user without any stored data; after this user interacts the system starts to observe the user, which could be considered as an imprecise projection of his preference relation. With the global and observed (user) relations it could be possible to determine the best fitting user preference models for the new user only by investigating the similarities. The same principles could be used to speed up the "warm up" period and to overcome the cold-start problem; the user preferences are determined according to the relation between the fuzzy relations expressed by his/her consumption patterns through feedback (or meta data etc) and the fuzzy relation representing the space of known user profiles.

Equivalence and/or similarity measure of fuzzy relations could be used to find the best fitting items with a specific (fuzzy relation) model in a global (fuzzy relation) space. It is easy to see that the computational requirements of the method could be decreased by limiting the calculation of similarity measure of the more specific fuzzy relation and the part of the global fuzzy relation, which is limited by the support of the more specific fuzzy relation. A simple illustration of fuzzy relations describing a user and global preferences can be seen in Fig. 2.

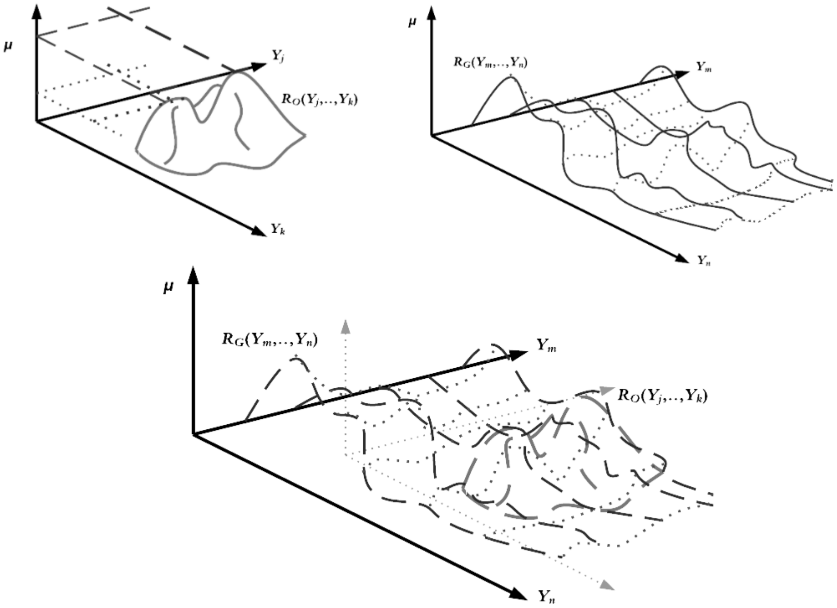


Fig. 2. Illustration of "specific" and "global" fuzzy relations describing items or users

Misclassified or purely detailed items without or with only a limited number of feedbacks are difficult to recommend, and the accuracy of these hugely depends on the quality of meta data. Some of the methods use meta data enrichment to overcome this problem. The projections and cylindrical extensions or closures of fuzzy relations representing the user preferences, feedbacks or even statistical parameters could be used to reduce the computational requirements of the methods by decreasing the complexity of the models, but it is also possible to apply these methods to fill out missing data or preferences of a single item or user by typical values of other similar preference relations with the same concepts shown above.

It is easy to see from the previous and this section that, there is a need for a comprehensive and detailed overview of fuzzy methods in recommender systems and it is reasonable to investigate these problems from mathematical aspects, but these are well over the limits of this work.

6 Summary and Future Work

In this paper the common problems of recommender systems were addressed from industrial point of view, basics of fuzzy methods and some of their applications were summarized. The mathematical backgrounds of these fuzzy relational concepts in recommender systems should be researched and developed in detail, implemented and compared with actual systems. In order to achieve the proposed object, the following main phases should be executed: (1) a representative detailed data set should be collected (on both users and items), (2) detailed mathematical research of the application of fuzzy relational calculus (might include the development of new operators and methods in the field), (3) design and implementation of a recommendation system based on the output of previous phases, (4) comparison of actual and the developed systems and (5) integrating and testing selected methods in real life environment.

Acknowledgement

This paper is partially supported by the GOP-1.1.1-11-2012-0172, the TÁMOP-4.2.2.A-11/1/KONV-2012-0012 and Hungarian Scientific Research Fund (OTKA) grants: K105529, K108405.

References

1. P. Baranyi and A. Csapo, "Definition and Synergies of Cognitive Infocommunications", Acta Polytechnica Hungarica, vol. 9, no. 1, pp. 67-83, 2012.
2. B. Szi and A. Csapo, "An outline of some human factors contributing to mathability research", In the Proceedings of 5th IEEE Conference on Cognitive Infocommunications, Vietri sul Mare, pp. 583-586, 2014.
3. G. Jawaheer, M. Szomszor, P. Kostkova: Comparison of implicit and explicit feedback from an online music recommendation service, In Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec '10). ACM, USA, pp. 47-51, 2010.
4. X. Amatriain, J. Pujol, N. Tintarev, N. Oliver: Rate it again: increasing recommendation accuracy by user rerating. Proceedings of the third ACM conference on Recommender systems, ACM, pp. 173–180, 2009.
5. S.S. Anand, P. Kearney, M. Shapcott: Generating semantically enriched user profiles for Web personalization. ACM Transactions on Internet Technology 7:4, p. 22, 2007.
6. Y. Hu, Y. Koren, C. Volinsky: Collaborative Filtering for Implicit Feedback Datasets, 2008 Eighth IEEE International Conference on Data Mining, pp. 263-272, 2008.
7. X. Amatriain, J. Pujol, N. Oliver: I like it... I like it not: Evaluating User Ratings Noise in Recommender Systems, User Modeling, Adaptation, and Personalization. Springer, Berlin, pp. 247-258, 2009.
8. D. Parra, A. Karatzoglou, X. Amatriain, I. Yavuz: Implicit feedback recommendation via implicit-to-explicit ordinal logistic regression mapping, Proceedings of the CARS-2011, USA, 5 pages, 2011.
9. R. Van Meteren and M. Van Someren, "Using content-based filtering for recommendation", In Proceedings of the Machine Learning in the New Information Age: MLnet/ECML2000 Workshop, pp. 47–56, 2000.
10. A. M. Rashid, G. Karypis, and J. Riedl. Learning Preferences of New Users in Recommender Systems : An Information Theoretic Approach. In ACM SIGKDD Explorations Newsletter, volume 10, page 90, Dec. 2008.
11. G. Shaw, Y. Xu, and S. Geva. Using Association Rules to Solve the Cold-Start Problem in Recommender Systems. In Advances in Knowledge Discovery and Data, pages 21-24, 2010.

12. H. Sobhanam, and a. K. Mariappan. Addressing cold start problem in recommender systems using association rules and clustering technique. In ICCCI, pp. 1-5, 2013.
13. R. Burke, "Hybrid recommender systems: Survey and experiments", In User Modeling and User-Adapted Interaction, 12(4), pp. 331–370, 2002.
14. J. Basilico, and T. Hofmann, Unifying collaborative and content-based filtering. In the Proceedings of The Twenty-First International Conference on Machine Learning, AXM p. 9, 2004.
15. L. A. Zadeh: Fuzzy Sets, in Information and Control, Vol. 8, pp. 338-353, 1965.
16. L.A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning I, II, III. Information Science, 8:199–251, 301–357, 9:43–80, 1975.
17. L.A. Zadeh. Fuzzy logic and approximate reasoning. Synthese, 30(1):407–428, 1975.
18. L.T. Kóczy, and M. Hajnal, "Cluster analysis in the caryometry applying a new fuzzy algebra ", In W.J. Perkins (ed.), Biomedical Computing, Pitmans Medical Publishing Co. Ltd., Tunbridge Wells, pp. 183-190, 1977.
19. L.T. Kóczy, "Vector valued fuzzy sets", BUSEFAL- BULL STUD EXCH FUZZIN APPL, pp. 41-57, 1980.
20. L.T. Kóczy, M. Hajnal, Classification of textures by vectorial fuzzy sets, In: M M Gupta, E Sanchez (ed.) Approximate Reasoning in Decision Analysis, Amsterdam; Oxford; New York; Tokyo: North-Holland Publishing Company, pp. 157-164, 1982
21. Cornelis, C., Guo, X., Lu, J., & Zhang, G. (2005). A Fuzzy Relational Approach to Event Recommendation. Proceedings of 2nd Indian International Conference on Artificial Intelligence (IICAI'05).
22. Al-Shamri, M. Y. H., & Bharadwaj, K. K. (2008). Fuzzy-genetic approach to recommender systems based on a novel hybrid user model. Expert Systems with Applications, 35(3), 1386–1399.
23. Kant, V., & Bharadwaj, K. K. (2012). Enhancing recommendation quality of content-based filtering through collaborative predictions and fuzzy similarity measures. Procedia Engineering, 38, 939–944.
24. Zhang, Z., Lin, H., Liu, K., Wu, D., Zhang, G., & Lu, J. (2013). A hybrid fuzzy-based personalized recommender system for telecom products/services. Information Sciences, 235, 117–129.
25. Son, L. H. (2014). HU-FCF: A hybrid user-based fuzzy collaborative filtering method in Recommender Systems. Expert Systems with Applications, 41(15), 6861–6870.
26. Perez, L. G., Barranco, M., & Martinez, L. (2007). Building user profiles for recommender systems from incomplete preference relations. In Fuzzy Systems Conference, 2007. FUZZ-IEEE 2007. IEEE International (pp. 1–6). IEEE.
27. Zenebe, A., Zhou, L., & Norcio, A. F. (2010). User preferences discovery using fuzzy models. Fuzzy Sets and Systems, 161(23), 3044–3063.
28. Porcel, C., & Herrera-Viedma, E. (2010). Dealing with incomplete information in a fuzzy linguistic recommender system to disseminate information in university digital libraries. Knowledge-Based Systems, 23(1), 32–39.
29. Nilashi, M., Ibrahim, O. bin, & Ithnin, N. (2014). Hybrid recommendation approaches for multi-criteria collaborative filtering. Expert Systems with Applications, 41(8), 3879–3900.
30. Nadi, S., Saraee, M., & Davarpanah-Jazi, M. (2010). A fuzzy recommender system for dynamic prediction of user's behavior. In Internet Technology and Secured Transactions (ICITST), 2010 International Conference for (pp. 1–5). IEEE.
31. Liu, W., & Gao, L. (2014). Recommendation System Based on Fuzzy Cognitive Map. Journal of Multimedia, 9(7), 970–976.

32. Birtolo, C., & Ronca, D. (2013). Advances in Clustering Collaborative Filtering by means of Fuzzy C-means and trust. Expert Systems with Applications, 40(17), 6997–7009.
33. Lu, J. (2004). Personalized e-learning material recommender system. In International conference on information technology for application (pp. 374–379).
34. Wu, D., Zhang, G., & Lu, J. (2013). A fuzzy tree similarity measure and its application in telecom product recommendation. In Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on (pp. 3483–3488). IEEE.
35. Castro-Schez, J. J., Miguel, R., Vallejo, D., & López-López, L. M. (2011). A highly adaptive recommender system based on fuzzy logic for B2C e-commerce portals. Expert Systems with Applications, 38(3), 2441–2454.
36. Ramkumar, V., Rajasekar, S., & Swamynathan, S. (2010). Scoring products from reviews through application of fuzzy techniques. Expert Systems with Applications, 37(10), 6862–6867.
37. Cornelis, C., Lu, J., Guo, X., & Zhang, G. (2007). One-and-only item recommendation with fuzzy logic techniques. Information Sciences, 177(22), 4906–4921.
38. García-Crespo, Á., López-Cuadrado, J. L., González-Carrasco, I., Colomo-Palacios, R., & Ruiz-Mezcua, B. (2012). SINVLIO: Using semantics and fuzzy logic to provide individual investment portfolio recommendations. Knowledge-Based Systems, 27(0), 103–118.
39. Terán, L., & Meier, A. (2010). A fuzzy recommender system for eelections. In Electronic Government and the Information Systems Perspective (pp. 62–76). Springer.
40. Dyczkowski, K., & Stachowiak, A. (2012). A recommender system with uncertainty on the example of political elections. In Advances in Computational Intelligence (pp. 441–449). Springer.

A method to analyse road risks using artificial vision and fuzzy logic

Juan Giralt, Juan Moreno-Garcia, Luis Jimenez-Linares, and Luis Rodriguez-Benitez

Information Systems and Technologies Department
University of Castilla-La Mancha (Spain)
{juan.giralt,juan.moreno,luis.jimenez,luis.rodriguez}@uclm.es
http://webpub.esi.uclm.es/investigacion/grupos/oreto

Abstract. This paper presents a system to detect abnormal movements of a vehicle in a road. This technique compares the trajectory of the vehicle with the information obtained from the lane marks of the road and detects lane changes and vehicle skids. Then, a process to obtain one single value representing the road's shape (left bend, straight-away, right bend) is done, to do this, each of the slopes of the two lines marking the edges of traffic lanes are computed. Then a comparison process to detect those frames where there is a logical correspondence between the vehicle displacement and the road shape is performed. The proposed system takes only as input information from the H264/AVC motion vectors and the videos are captured from a moving vehicle. As output the time intervals in which the vehicle displacement corresponds with a risky situation are obtained.

Keywords: motion vectors, H264/AVC, lane mark segmentation, linguistic comparison

1 Introduction

Intelligent Transport Systems use advanced technologies to improve vehicle's safety, for example, systems using computer vision techniques to segment the lane marks in a road [6], perform monitoring [8], and alerting to deviations in the path of the vehicle detected from information highway lines [1, 4]. There are also works dealing with this problem using fuzzy logic, such as Wang et al. [7] and Obradovic et al. [5]. One of the main features of the proposed method is its efficiency, that is because it works with very little input information. This is possible by taking as input data the motion vectors of the video compression standard H264/AVC. This standard uses motion compensation based on macroblocks, pixel arrays, etc. and exploits duplicate information in successive frames. Similar information present in a frame and another, called reference frame, are not stored, only a motion vector representing the displacement between macroblocks is used to code the macroblock spatial differences. The motion vector field obtained by H264/AVC can be considered as the sparse and imprecise approximation of the optical flow.

The rest of the paper is organized as follows. Section 2 presents the mechanism that allows to represent by means of a fuzzy value the information obtained from the segmentation of the lane marks of the road. Later, in Section 3 an algorithm to compare the vehicle displacement with the road shape is detailed. Finally, in Sections 4 and 5 the experimental results, the conclusions and the future works are shown.

2 Road's shape detection and representation

Algorithm 1 shows how to obtain a list containing the road's shape changes all along the video. The first step of this algorithm is based on a technique presented in [2]. This method detects the lane marks of the road that are represented using a set of statistical attributes. In this paper only is needed the information from the left and right lane marks slopes named $s(LL)$ and $s(RL)$, respectively.

Algorithm 1 Computing the list L'

```
1: $RS \leftarrow$ detection of line geometry using [2]
2: $FRSList \leftarrow$ fuzzification of each $RS$ and computation of the list
3: $L \leftarrow$ Grouping consecutive elements of $FRSList$
4: $L' \leftarrow$ Grouping consecutive elements of $L$
```

Figure 1 shows the values of the left and right slopes of the first 1000 frames of a sample video. The road turns left when the line slopes decrease, and the road turns right when the slopes increase their values. If it is not possible to detect a line in a concrete frame the method assumes the last value detected as it happens in sharp turns, for example, in roundabouts [2].

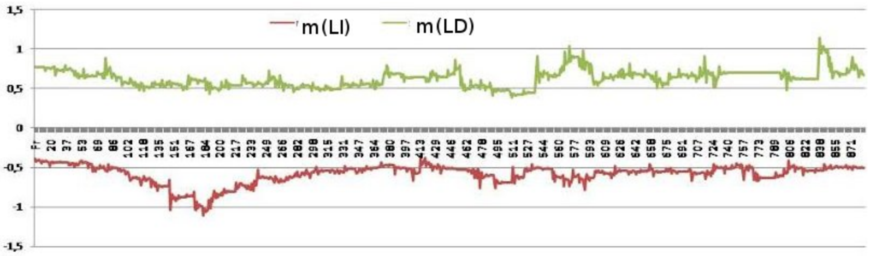


Fig. 1. Slopes of the left and right lines of the lane

A single measure called $s(m)$ is calculated to automatically obtain the road shape combining the slopes $s(LL)$ and $s(RL)$ for every frame (Equation 1).

$$s(m) \leftarrow s(LL) + s(RL) \tag{1}$$

Equation 2 assigns a description in natural language to the road's shape in every frame of the video. Two threshold values are needed (U_{min} and U_{max}). These values are obtained empirically based on the minimal and maximal values of $s(m)$ in a frame sequence while the vehicle is moving in a straight line. This sequence could be considered as a "training sequence". U_{min} and U_{max} are -0.1 and 0.18 in Figure 2. The list containing the representative values of each frame is called *Road Shape* (RS).

$$text \leftarrow \begin{cases} \textit{Left bend}, \ If\ s(m) < U_{min} \\ \textit{Straight}, \quad If\ U_{min} \leq s(m) \leq U_{max} \\ \textit{Right bend}, If\ s(m) > Umax \end{cases} \tag{2}$$

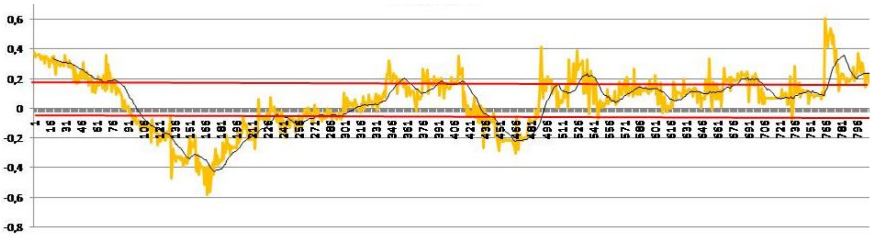


Fig. 2. Joint slopes and threshold values.

A linguistic variable [9] called *Road Shape* (RSV) is used now. The reason to translate the representation to the fuzzy domain is to manage the inherent noise in the H264 motion vectors and the use of a mobile camera that also add noise to the captured data (step 2 of Algorithm 1). This variable allows to add fuzzy boundaries around U_{min} and U_{max}, and it is composed of three trapezoidal fuzzy sets corresponding to the linguistic labels *Left Bend* ($LB = \{-0.5, -0.5, U_{min} - k, U_{min} + k\}$), *Right Bend* ($RB = \{U_{min} - k, U_{min} + k, U_{max} - k, U_{max} + k\}$) and *Straight* ($S = \{U_{max} - k, U_{max} + k, 0.5, 0.5\}$) respectively. The parameter k is determined empirically and it is based on the maximum and minimum values of $s(m)$ in each concrete video. RSV support is defined between $[-0.5, 0.5]$ for the videos used in the experimentation.

The fuzzification process consists of the generation of tuples named $RSFuzzy$ in every frame and their structure is shown in Equation 3.

$$RSFuzzy = (frame, \mu_{LC}(s(m)), \mu_S(s(m)), \mu_{RC}(s(m))) \tag{3}$$

Each one of these tuples are stored in a list ordered by the number of frame (Equation 4).

$$FRSList \leftarrow FRSList + \{RSFuzzy\} \tag{4}$$

After that (step 3 of Algorithm 1), $SetSize$ consecutive elements are processed to obtain a single value that represents these elements. $SetSize$ is empirically defined, and it must be proportional to the modulus of the motion vectors. The more speed is reached the less the value of $SetSize$ is, because more distance is covered in less time and there can be more rapid changes in the shape of the road. Now, a linguistic value is going to represent the road's shape in $SetSize$ consecutive elements of $FRSList$. This value is named *Label* and it is obtained as the linguistic variable with maximum membership once all the memberships to the same variable are added for the $SetSize$ frames. Then, in Equation 5 the new linguistic representation for $SetSize$ frames is shown.

$$MaxFCFuzzy = (frame\_ini,\ frame\_fin,\ Label) \tag{5}$$

Finally, in the last step of the Algorithm 1, a new clustering process is done and it obtains L'. This process is needed to joint consecutive elements in L with the same value for the attribute *Label*. Equation 6 shows the structure of the tuple containing the union of consecutive elements in L satisfying this condition.

$$MaxFCFuzzy = (frame\_ini,\ frame\_fin + SetSize,\ Label) \tag{6}$$

In brief, L' is a list that contains the road's shape through the time using $MaxFCFuzzy$ tuples to represent it. For example, Equation 7 details how a video of 3420 frames is represented.

$$L' = \{(0,\ 139,\ S),\ (140\ ,389\ ,RB)...,(3300\ ,3419\ ,S)\} \tag{7}$$

3 Detection of risky situations

The detection of risky situations is done by means of a comparison process between the information contained in L' and the results obtained in [3] about the characterization of the vehicle displacement. This method is similar to the one proposed in Section 2. It generates a tuple called *Video Vehicle Displacement* (VVD) with a similar structure to L', where the linguistic labels used to represent the displacement are *Turning Left* (TL), *Straight* (S) and *Turning Right* (TR). Table 1 shows an example of the elements to be compared. The time intervals are different for each one of the lists since they are obtained using different processes. Then a mechanism to establish common time intervals that represent the same interval times must be done. This process is described in Section 3.1, after that, the comparison process is detailed (Section 3.2).

3.1 Obtaining the common intervals

Two ordered lists are obtained to identify the common intervals. These lists contain the initial frame of the video and the attributes that define the final of each temporal interval in each one of the lists. This attribute is named $frame\_fin$ in

Table 1. L' (left) and VVD (right) to compare.

f_ini	f_fin	**Label**	f_ini	f_fin	**Label**
0	109	S	0	119	S
110	283	LB	120	270	TL
284	327	S	271	329	S
328	458	RB	330	470	TR
459	479	S	471	482	S
481	490	LB	483	509	TL

Equation 5. Equations 8 and 9 show the two lists from the information shown in Table 1.

$$L_1 \leftarrow \{0, 109, 283, 327, 458, 479, 490\} \tag{8}$$

$$L_2 \leftarrow \{0, 119, 270, 329, 470, 482, 509\} \tag{9}$$

After that, the two lists are merged in order to obtain an ordered set named *Common Interval* (CI). As mathematical sets, it does not allow duplicate elements. For example, from the lists of the Equations 8 and 9 the $CI = \{0, 109, 119, 270, 283, 327, 329, 458, 470, 479, 482, 490, 509\}$ is obtained. Each two consecutive values of CI is now a new interval. Both L' and VVD are reorganized using these intervals as it is shown in Table 2.

Table 2. New temporal values in L' (left) and VVD (right).

f_ini	f_fin	**Etiq.**	f_ini	f_fin	**Etiq.**
0	109	S	0	109	S
110	119	LB	110	119	S
120	270	LB	120	270	TL
271	283	LB	271	283	S
284	327	S	284	327	S

3.2 Comparison process

The elements of L' and VVD are now compared in the same temporal interval in order to detect the correspondences between the lane marks and the displacement of the vehicle. More concretely, the attribute *Label* of L' and VVD is used and from Equation 10 risky situations can be obtained. For example, this equation returns *TRUE* for the interval $[0, 109]$ and *FALSE* for the interval $[110, 119]$ (Table 2).

$$Comparison \leftarrow \begin{cases} True, & Label(L') = LB \; and \; Label(VVD) = TL \\ True, & Label(L') = S \; and \; Label(VVD) = S \\ True, & Label(L') = RB \; and \; Label(VVD) = TR \\ False, & In \; other \; case \end{cases} \tag{10}$$

The output of the comparison process is stored in a List of Differences (LD) containing the tuples with the following structure: $\{frame\_ini$, $frame\_fin$, $Label_{L'}$, $Label_{VVD}\}$. These are the intervals with discrepancies between lane marks and the displacement of the vehicle.

4 Experimental results

In the experimentation, three videos are used. One is a video recorded by the authors whilst the two other videos were captured in a World Rally Car competition. The two drivers were Sebastian Loeb and Peter Solberg. We consider these two last videos very difficult for the aims of these tests since there are sudden changes in the direction in the speed and there a lot of vegetation in the ditch and in the road edge because the car drives along a rural minor road. This videos are identified as *Own*, *Loeb* y *Solberg*, respectively.

4.1 Evaluation of the detection of the road's shape

Table 3 exposes the obtained results for each one of the experiments in the process of road's shape recognition. The success rate is considerably lower for Solberg and Loeb tests and failures mainly occur when the car begins a sharp curve with a high speed. That is because the lane mark segmentation algorithm loses the line followed and this error affects to the high level recognition process here detailed.

Table 3. Road characterization.

	Own	**Loeb**	**Solberg**
Hits	89.6%	67.7%	60.2%
Errors	10.4%	32.3%	38.8%

4.2 Evaluation of risky situations

Table 4 shows the comparison process for one of the experiments. When *False* is obtained as result, a new element to the list of differences is added.

Then, hits of the system occurs when an element in the list of differences corresponds with a risky situation in the video. The errors occur when the proposed system does not detect risky situations as it is shown in Table 5 (from the

Table 4. Comparison results in the experiment *Own*

f_ini	f_fin	**VVD**	**L'**	**Comparison**
0	26	S	S	True
27	47	S	RB	False
48	113	S	S	True
114	244	TL	LB	True
245	283	TL	S	False
284	327	S	S	True
328	375	TR	S	False
376	397	TR	RB	True
398	451	TR	S	False
452	462	TR	RB	True
463	495	S	S	True
496	527	S	LB	False
528	571	S	S	True
572	593	S	RB	False
594	710	S	S	True
711	752	TR	S	False
753	796	TL	S	False
797	807	S	S	True
808	839	TR	S	False
840	909	S	S	True
910	942	S	LB	False
943	999	S	S	True

column 1 to 4) where risky situations are not present in the list of differences. The percentage of the table refers to the number of frames. In *Loeb* experiment, our system reaches a 97.7% of hits, then the 2.3% corresponds to risky situations not detected.

Table 5. Risky detection in frames not present in *LD*

frames not present in *LD*				**frames present in *LD***			
	Own	**Loeb**	**Solberg**		**Own**	**Loeb**	**Solberg**
Hits	100%	97.7%	76.3%	**Hits**	32.5%	30.2%	29.9%
Errors	0%	2.3%	23.7%	**Errors**	67.5%	69.8%	70.1%

Table 5 (from the column 5 to 8) details the results of the hits of our system. A 29.9% detects the risky situations in the experiment *Solberg* (70.1% can be considered errors). This percentage of errors are mainly due to the difficulty of this video and the fact that the car continuously drives from one lane to the another and other unexpected behaviours.

5 Conclusions and future works

In this work we have presented a new technique to represent in a linguistic way the road shape from a video sequence using as input data the H264/AVC motion vectors. The use of fuzzy logic allows to work with linguistic representations and to process the information from several frames simultaneously. The use of a very little amount of data allows to obtain a no time-consuming method. The use of linguistic variables to represent the vehicle displacement and the road shape makes interpretable the comparison process proposed. Despite the difficulty of the selected videos for the experimentation, acceptable results have been obtained in this first approach. As future works, it should be developed mechanisms to analyse in more detail each one of the differences obtained as output of the comparison process. Issues as their duration or the degree of differentiation between vehicle displacements and road's shape are factors that should be taken into account for future developments.

Acknowledgement

This work has been funded by the Information Systems and Technologies Department of the University of Castilla-la Mancha.

References

1. A. Borkar, M. Hayes, and M. T. Smith. A new multi-camera approach for lane departure warning. In *Proceedings of the 13th international conference on Advanced concepts for intelligent vision systems*, Springer-Verlag, Berlin, Heidelberg, pp. 58–69. 2011.
2. J. Giralt, L. Rodriguez-Benitez, J. Moreno-Garcia, C.J. Solana-Cipres, L. Jimenez: Lane mark segmentation and identification using statistical criteria on compressed video. *Integrated Computer-Aided Engineering* 20, 2, pp. 143–155, 2013.
3. J. Giralt, J. Moreno-Garcia, L. Jimenez-Linares L., E. Del Castillo, L. Rodriguez-Bentez. A Fuzzy Representation of Vehicle Trajectories using Motion Data from H264/AVC Video. In *Proceedings of the 14th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2014*, J. Vigo-Aguiar Ed., 2, pp. 635–646, 2014.
4. J. Lee and U. Yi. A lane-departure identification based on LBPE, Hough transform, and linear regression. *Comput. Vis. Image Underst* 99, 3 , pp. 359–383, 2005.
5. I. Obradovic, Z. Konjovic, E. Pap, and I. J. Rudas. Linear fuzzy space based road lane model and detection. *Know.-Based Syst*, pp. 37–47. 2013.
6. T. Tran, J. Son, B. Uk, J. Lee, and H. Cho. An adaptive method for detecting lane boundary in night scene. In *Proceedings of the Advanced intelligent computing theories and applications.* Springer-Verlag, Berlin, Heidelberg, pp. 301–308. 2010.
7. J. Wang, C. Lin, and S. Chen. Applying fuzzy method to vision-based lane detection and departure warning system. *Expert Syst. Appl.*, 37, 1, pp. 113–126. 2010.
8. Y. Wang, N. Dahnoun, and A. Achim. A novel system for robust lane detection and tracking. *Signal Process.*, 92, pp. 319–334. 2012.
9. L. A. Zadeh: Similarity relations and fuzzy orderings. *Information Sciences*, pp. 177–200. 1971.

Using genetic algorithms to learn a fuzzy based pseudometric for k-NN classification

F. Martins[1], J. C. Becceneri[1], L. Dutra[1], D. Lu[2], and S. Sandri[1]

[1] Instituto Nacional de Pesquisas Espaciais
12201-970, São José dos Campos, SP, Brazil
E-mail: flavinha@dpi.inpe.br, becce@lac.inpe.br, luciano.dutra@dpi.inpe.br,
sandra.sandri@inpe.br
[2] Michigan State U. - Center for Global Change and Earth Observations, East Lansing, MI, USA
E-mail: ludengsh@msu.edu

Abstract. We address the derivation of pseudometric based on fuzzy relations for classification applications, by the use of genetic algorithms to learn the fuzzy relations. We present an experiment for the classification of land use in an area of the Brazilian Amazon region.

Keywords: k-NN classification, fuzzy partitions, genetic algorithms

1 Introduction

In a previous work [2], we proposed a a function called f^+, based on fuzzy relations, which are themselves derived from fuzzy partitions, for use in classification applications. This function is the complement in $[0,1]$ of a particular kind of fuzzy relation, called an Order Compatible Fuzzy Relation (OCFR$_\preceq$), defined using a total order $(\Omega, \preceq)$ [10]. An OCFR$_\preceq$ itself is derived from a type of fuzzy partition (a collection of fuzzy sets), called Convex Fuzzy Partitions (CFP$_\preceq$). The creation of OCFR$_\preceq$ was motivated by the need to ease the burden of creating suitable relations for use the particular fuzzy case-based reasoning classification approach proposed in [8]. In [2], we proved that f^+ function is i) a pseudometric, when obtained from a specific type of CFP$_\preceq$, called 2-Ruspini, and, in particular, a ii) metric, when this CFP$_\preceq$ is moreover composed solely of triangular fuzzy sets. The same happens in the case of multidimensional domains, for function $f^+_{(\mu)}$ that aggregates the results obtained for f^+ in each domain, using the arithmetic means as aggregation operator μ.

Here we address the derivation of f^+ for k-NN classification applications [11], by the use of fuzzy genetic algorithms [1] to learn the fuzzy relations. We describe an application in the classification of land cover and use in an area of the Brazilian Amazon region.

2 Fuzzy relation based pseudometrics f^+ and $f^+_{(\mu)}$

Let $S : \Omega^2 \to [0,1]$ be a fuzzy binary relation and $(\Omega, \preceq)$ be a total order. Formally, S is an Order Compatible Fuzzy Relation with Respect to a Total Order $(\Omega, \preceq)$ (*OCFR*$_\preceq$ or *OCFR*, for short), when it obeys the following properties [10]:

- $\forall x, y, z \in \Omega$, $S(x,x) = 1$ (*reflexivity*)
- $\forall x, y, z \in \Omega$, $S(x,y) = S(y,x)$ (*symmetry*)
- $\forall x, y, z \in \Omega$, if $x \preceq y \preceq z$, then $S(x,z) \leq \min(S(x,y), S(y,z))$ (*compatibility with total order* $(\Omega, \preceq)$, or $\preceq$*-compatibility* for short).

Let $(\Omega, \preceq)$ be a total order and let $\mathbf{A} = \{A_1, ..., A_t\}$ be fuzzy partition (a collection of fuzzy sets) in Ω; here A_i denotes a fuzzy set but also its associated membership function. Let the core and support of a fuzzy set A be defined as $core(A) = \{x \in \Omega \mid A(x) = 1\}$ and $supp(A) = \{x \in \Omega \mid A(x) > 0\}$), respectively [3]. Formally, $\mathbf{A}$ is a Convex Fuzzy Partition with Respect to a Total Order $(\Omega, \preceq)$ (*CFP*$_\preceq$ or *CFP*, for short), if it obeys the following properties [10]:

1. $\forall A_i \in \mathbf{A}, \exists x \in \Omega, A_i(x) = 1$ (*normalization*),
2. $\forall x, y, z \in \Omega, \forall A_i \in \mathbf{A}$, if $x \preceq y \preceq z$ then $A_i(y) \geq \min(A_i(x), A_i(z))$ (*convexity*),
3. $\forall x \in \Omega, \exists A_i \in \mathbf{A}, A_i(x) > 0$ (*domain-covering*),
4. $\forall A_i, A_j \in \mathbf{A}$, if $i \neq j$ then $core(A_i) \cap core(A_j) = \emptyset$ (*non-core-intersection*).

Let $\mathcal{A}_{(\Omega,\preceq)}$ denote the set of all CFPs that can be derived considering a total order $(\Omega, \preceq)$. CFP $\mathbf{A} \in \mathcal{A}_{(\Omega,\preceq)}$ is said to be a n*-CFP* if each element in Ω has non-null membership to at most n fuzzy sets in $\mathbf{A}$ ($n \geq 1$). In particular, a 2-CFP$_\preceq$ $\mathbf{A}$ is called a *2-Ruspini* partition, when it obeys additivity:

- $\forall x \in \Omega$, $\sum_i A_i(x) = 1$ (*additivity*)

In [10], the authors propose to generate OCFR$_\preceq$ $S^+ : \Omega^2 \rightarrow [0,1]$ from a CFP$_\preceq$ $\mathbf{A}$ as

$$S^+(x,y) = \begin{cases} 0, \text{if } S^*(x,y) = 0 \\ S_L(x,y), \text{otherwise} \end{cases}$$

$$\forall x, y \in \Omega, S^*(x,y) = \sup_i \min(A_i(x), A_i(y))$$

$$\forall x, y \in \Omega, S_L(x,y) = \inf_i \quad 1- \mid A_i(x) - A_i(y) \mid$$

Note that S_L is constructed based on the Lukasiewicz biresiduated operator [9].

In [2], the following function was proposed for tasks in which metrics and pseudometrics are employed[1]:

$$\forall x, y \in \Omega, f^+_{\mathbf{A}}(x,y) = 1 - S^+_{\mathbf{A}}(x,y).$$

This formula can be written directly as:

$$\forall x, y \in \Omega, f^+_{\mathbf{A}}(x,y) = \begin{cases} 1, \text{if } \forall i, \min(A_i(x), A_i(y)) = 0, \\ \sup_i \quad \mid A_i(x) - A_i(y) \mid, \text{otherwise.} \end{cases}$$

[1] A metric satisfies non-negativity, symmetry and the triangle inequality and the identity of indiscernibles properties. Pseudometrics obey the same properties, except for the identity of indiscernibles, that is substituted by anti-reflexivity, a weaker property.

When no confusion is possible, we denote $f^+_{\mathbf{A}}$ as simply f^+.

Let $O = \Omega_1 \times ... \times \Omega_m$, where $\forall i, (\Omega_i, \preceq)$ is a total order. Let $\mathbf{A}_i$ be a 2-Ruspini CFP$_\preceq$ on Ω_i and f^+_i be derived from $\mathbf{A}_i$. Let $f^+_{(\mu)} : O \rightarrow [0,1]$ be the extension of function f^+ to multidimensional domains, defined as

$$f^+_{(\mu)}(x, y) = \mu(f^+_1(x_1, y_1), ..., f^+_m(x_m, y_m)),$$

where $\mu : [0,1]^m \rightarrow [0,1]$ is the arithmetic mean, i.e., $\mu(a_1, ..., a_m) = \frac{\sum_{1 \leq i \leq m} a_i}{m}$.

In [2], it is proved that $f^+_{\mathbf{A}}$ is a pseudometric, in general, and a distance when all fuzzy sets in **A** are triangular. Function $f^+_{(\mu)}$ trivially satisfies symmetry, anti-reflexitivity and non-negativity. The same result holds for $f^+_{(\mu)}$. In the same work, function $f^+_{(\mu)}$ was tested in a real-world application and yielded very good results when compared to both the Euclidean and Mahalanobis distances.

3 Learning $f^+_{(\mu)}$ using genetic algorithms for k-NN classification

We propose to use genetic algorithms to learn the fuzzy partitions necessary for function $f^+_{(\mu)}$, which is also our fitness function. Here we consider classification by k-NN but other methods could be used.

Let $X = \{x_1, ...x_m\}$ be a set of variables, each of which defined in domain $\Omega_i = [l_i, u_i], i \in \{1, m\}$. We encode each chromosome as a sequence of m genes, each of which related to a variable in X. The i-th gene is a sequence of parameters $< p_1, ..., p_s >$, representing points in domain Ω_i for a Ruspini partition. The sequence is such that $p_i \leq p_{i+1}, 1 \leq i \leq s - 1$. In a trapezoidal partition, the first (respec. last) fuzzy term will have $[l_i, p_1]$ (respec. $[p_s, u_i]$) as core and $[l_i, p_2]$ (respec. $[p_{s-1}, u_i]$) as support. In a triangular partition, the first (respec. last) fuzzy term will have l_i (respec. u_i) as core and $[l_i, p_1]$ (respec. $[p_s, u_i]$) as support.

Crossover consists in choosing a cutting place in two selected chromosomes c_1 and c_2, and generating two new chromosomes c_{12} and c_{21}. Let chromosome c_i be described as $< p_{i,1}, ..., p_{i,s} >$ and let the cutting happen between the (k)-th and (k+1)-th genes. The crossover between any two chromosomes c_1 and c_2 would be generate two new chromosomes c_{12} and c_{21}, respectively described as $< p_{1,1}, ..., p_{1,k}, p_{2,k+1}, , ..., p_{2,s} >$ and $< p_{2,1}, ..., p_{2,k}, p_{1,k+1}, , ..., p_{1,s} >$

If one of the generated chromosomes does not satisfy the condition on the p_is, we reorganize the parameters. For example, let us suppose we have two chromosomes with 3 trapezoidal fuzzy sets Let c_1 and c_2 be described as $< 10, 20, 30, 40 >$ and $< 31, 32, 33, 34 >$, respectively, and that the cutting point is between p_2 and p_3. We obtain a valid chromosome, $c_{12} =< 10, 20, 33, 34 >$, and an invalid one, $c_{21} =< 31, 32, 30, 40 >$. We then rearrange the invalid chromosome as $c_{21} =< 30, 31, 32, 40 >$.

In this work we use n-fold cross-validation. First of all, a data set T is partitioned in n (approximately) equal parts (folds) T_i, such that $T = \cup_i T_i$. Then, for a given fold i, training is performed using the elements of all folds, except for those in i, and testing is performed the elements of fold i itself, making $Train_i = \bigcup_{T_j \in T, j \neq i} T_j$, and $Test_i = T_i$.

4 Experiments

In the following, we briefly describe an experiment that illustrates the use of function $f^+_{(\mu)}$ in a land use classification task in the Brazilian Amazon region. The area of interest covers approximately 411 km^2 and in the municipality of Belterra, state of Pará, in the Brazilian Amazon region, partially contained in the National Forest of Tapajós. An intense occupation process occurred in the region along the BR-163 highway (Cuiabá-Santarém), with opening of roads to establish small farms, after deforestation of primary forest areas [4]. As a result, there are mosaics of secondary vegetation in various stages, with pastures and cultivated areas embedded in a forest matrix [5].

In this application, 14 attributes have been considered, derived from either radar or optical satellite images, with 6 classes: forest, initial or intermediate regeneration, advanced regeneration or degraded forest, cultivated area, exposed soil, and pasture. The samples consist of 138 ground information based hand-made polygons. The attribute value for each polygon is the average of the values for the pixels composing it. The experiments have been done using 6 folds (5 for training and 1 for testing).

To obtain the lower (respec. upper) bound for a variable domain, we took the smallest (respec. largest) value from the elements in the fold, less (respec. plus) 20%. We have tested two types of partition for each variable, a triangular and a trapezoidal one, each of which with 3 fuzzy terms. In the triangular experiment, each partition is described by $< p_1 >$, where p_1 is the core of the middle triangular fuzzy term. In the trapezoidal experiment, each partition is described by $< p_1, p_2, p_3, p_4 >$, where $[p_2, p_3]$ is the core of the middle trapezoidal fuzzy term.

In our experiments, for each fold, the candidate population has 10 chromosomes. Each chromosome has 3 genes, each of which describing a partition corresponding to one of 3 variables used here. We have used an elitist genetic algorithm, keeping the best 6 elements and combining the 3 first elements to generate the new candidates that replace the worst 4 elements. We used a mutation rate of .2 and 400 generations.

We have used two kinds of population in the initial generation for each fold: "random" and "selected". In the selected first population for the fuzzy terms, the points are obtained from a fixed set of percentage vectors. Considering all domains to be normalized to [0,1], the selected population for the trapezoidal fuzzy sets corresponds to the set of 10 quadruples $< .20, .40, .60, .80 >$, $< .05, .28, .52, .76 >$, $< .23, .47, .71, .95 >$, $< .23, .47, .52, .76 >$, $< .23, .28, .52, .76 >$, $< .23, .47, .71, .76 >$, $< .4, .55, .7, .85 >$, $< .15, .55, .7, .85 >$, $< .15, .3, .7, .85 >$ and $< .15, .3, .45, .85 >$. The selected population for the triangular fuzzy sets is obtained by taking the arithmetic means between p_2 and p_3 from the trapezoidal fuzzy terms. It corresponds to $< .50 >$, $< .40 >$, $< .59 >$, $< .49 >$, $< .40 >$, $< .59 >$, $< .62 >$, $< .62 >$, $< .50 >$ and $< .37 >$.

Figure 4 brings the accuracy results for this application, considering k-NN with 1 to 6 neighbours, using the several versions of function $f^+_{(\mu)}$: trapezoid-based and triangle-based, considering selected and random initial populations (kNN_dFtz_s, kNN_dFtz_r, kNN_dFtg_s, kNN_dFtg_r). For comparison, the figure also brings the Euclidean distance (kNN_dE).

We see from the figures that all methods had high accuracy and that the best average results in the 6 folds were obtained with the use of f^+_{μ} for the triangular partitions. The best individual results, considering all folds, were the same methods for 1, 2 and 3

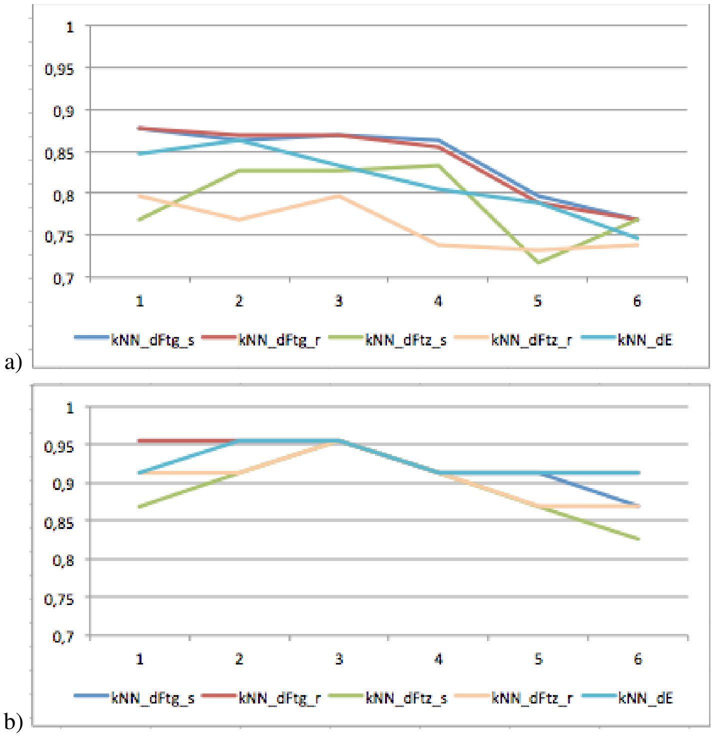


Fig. 1. Classification accuracy results for: a) k-NN average and b) k-NN maximum.

neighbours and the Euclidean distance for 2 and 3 neighbours. In particular, f^+_μ for the triangular partitions with the initial population obtained at random yielded the same results for the maximum as the Euclidean distance, except for 1 neighbour, when f^+_μ fares better. All methods fare better with a small number of neighbours. In particular, the best results for the triangular partitions, considering both the average and the maximum, is obtained already with a single neighbour. The worst results have been obtained with the trapezoidal partitions, for both types of initial populations.

5 Conclusions

In this work, we have proposed to use of genetic algorithms to learn fuzzy relations, that are parameters for a pseudometric $f^+_{(\mu)}$. We describe a classification application of land cover and use in an area of the Brazilian Amazon region, using k-NN. The results have shown that the triangular partitions produced the best results.

Future work includes experimenting with other data sets. We also intend to verify alternatives to reduce the computational cost, without a decrease in accuracy or adequately reducing the training data Another alternative consists in learning the partition for each variable separately; in order to calculate accuracy the distance relative to the other variables would be fixed (e.g. Euclidean) and aggregated with the distance obtained from the partition.

This work is a first step towards using $f^{+}_{(\mu)}$ in [7], an extension to k-NN for image classification, in which there is the possibility of using multiple spaces, that can be originated from different data sources, having different ranges of values, as well as the geographical space itself, allowing the use of topological associations.

Acknowledgements The authors acknowledge financial support from Brazilian research funding agencies CAPES, CNPq and FAPESP.

References

1. F. Herrera, Genetic Fuzzy Systems: Status, Critical Considerations and Future Directions, International Journal of Computational Intelligence Research, Vol.1, No.1, pp. 59-67 (2005)
2. Sandri, S., Martins-Bedê, F.T., Dutra, L.,: Using a Fuzzy Based Pseudometric in Classification. Proc. 14th Int. Conf. on Info. Proc. and Management of Uncertainty in Knowledge-Based Systems, IPMU 2014, Montpellier-Fr (2014)
3. Dubois, D., Prade, H.: Possibility Theory: An Approach to Computerized Processing of Uncertainty. Plenum Press, New York (1988)
4. Brazilian Institute of Environment and Renewable Natural Resources (IBAMA): Floresta Nacional do Tapajós Plano de Manejo. Vol I. (in Portuguese) Available at: <http://www.icmbio.gov.br/portal/images/stories/imgs-unidades-coservacao/flona_tapajoss.pdf>. Date accessed: 30 Dec. 2013 (2009)
5. Escada, M.I.S., Amaral, S., Rennó, C.D., Pinheiro, T.F.: Levantamento do uso e cobertura da terra e da rede de infraestrutura no distrito florestal da BR- 163. Repport INPE-15739-RPQ/824, INPE, S.J.Campos, Brazil. Available at: <http://urlib.net/sid.inpe.br/mtc-m18@80/2009/04.24.14.45>. Date accessed: 30 Dec. 2013 (2009)
6. Korsrilabutr, T., Kijsirikul, B.: Pseudometrics for Nearest Neighbor Classification of Time Series Data. Engineering Journal, Thailand, 13, May. 2009. Available at: <http://engj.org/index.php/ej/article/view/46>. Date accessed: 30 Dec. 2013 (2009)
7. Martins-Bedê, F.T.: Souza Reis, M., Pantaleão, E., Dutra, L., Sandri, S. An application of multiple space nearest neighbor classifier in land cover classification Proc. IGARSS'14, pp 1713–1716 (2014)
8. Mendonça, J.H., Sandri, S., Martins-Bedê, F.T., Guimarães, R., Carvalho, O.: Training strategies for a fuzzy CBR cluster-based approach, Mathware & Soft Computing, v. 20, pp 42–49 (2013)
9. Recasens, J.: Indistinguishability operators, modelling fuzzy equalities and fuzzy equivalence relations. Series: Studies in Fuzziness and Soft Computing, vol 260, Springer Verlag (2011)
10. Sandri, S., Martins-Bedê, F.T.: A method for deriving order compatible fuzzy relations from convex fuzzy partitions. Fuzzy Sets and Systems, pp 91–103 (2014)
11. Theodoridis, S. and Koutroumbas, K.: Pattern recognition, Academic Press, 3rd edition (2006)

Reducing information systems considering similarity relations

María José Benítez[1], Jesús Medina[1]*, Dominik Ślęzak[1]

[1]Departamento de Matemáticas Universidad de Cádiz, Spain
Email: mariajose.benitez, jesus.medina@uca.es
[2]Institute of Mathematics, University of Warsaw. Poland
Email: slezak@mimuw.edu.pl

Abstract. Attribute reduction is an important step in order to decrease the computational complexity to deriving information from databases. In this paper, we extend the notions of reducts and bireducts introduced in rough sets theory for attribute reduction purposes and let them work with similarity relations defined on attributes values. Hence, the related mathematical concepts will be introduced and the characterizations of the new reducts and bireducts will be given in terms of the corresponding generalizations of the discernibility function.

1 Introduction

Fuzzy Set Theory (FST) introduced by Zadeh [**?**] and Rough Set Theory (RST) proposed by Pawlak [7], are complementary approaches to treating imperfect knowledge: meanwhile the first one considers a certain degree of truth given, in the second one the available information is incomplete. Specifically, in the absence of exact information about a set, it is represented by a pair of sets, which are the lower approximation and the upper approximation of the set.

Although in the original version proposed by Pawlak, the considered approximations were classical sets, there have been introduced some new variants in which the approximations could be fuzzy sets. A first definition, the rough fuzzy sets, was given by Fariñas del Cerro and Prade in the eighties [3].

A very important part is to reduce the size of the database, without losing information or elements of judgment. To this end various types of so called reducts were presented and studied in the RST-related literature [1, 4, 6]

Bireducts extend classical RST-based notions of reducts in order to provide more flexibility in operating with subsets of attributes and subsets of objects that those attributes can efficiently describe [5, 9, 10]. The main objective of the bireducts is to reduce the original system preventing the occurrence of incompatibilities and eliminating existing noise in the original data.

In this paper we study representations of bireducts both in the classical case and in situations when the notion of equality is weakened towards similarity.

[*] Corresponding author.

The organization of the paper is the following: Some basics concepts related to the notion of similarity relation, the notions of δ-similar and δ-discordant are called in Section 2. Section 3 presents the basic definitions with Boolean in the new similarity enviroment. Conclusions and propects for future work are given in Section 4.

2 Preliminaries

In this paper the classical theory of propositional logic will be considered in order to interpret the expression of the discernibility function.

Definition 1. *A WFF is said to be in* disjunctive normal form *(DNF) if it is $\top$, $\bot$, a cube or a disjunction (possibly empty) of cubes.*

A WFF is said to be in conjunctive normal form *(CNF) if it is $\top$, $\bot$, a clause or conjunction (possibly empty) of clauses.*

The above normal forms may be reduced using absorption laws until none of them can be further reduced, obtaining the reduced forms:

Definition 2. *A DNF is said to be* restricted *(briefly, RDNF), if it satisfies that any cube contains a literal or its complementary and it does not contain repeated literals, and other cubes.*

A CNF is said to be restricted *(briefly, RCNF), if it satisfies that any clause contains a literal or its complementary literal and it does not contain repeated literals, and other clauses.*

The previous definitions are critical to introducing and managing discernibility function used in RST and will be generalized in this work to consider similarity relations. Now, we will recall the basic definitions of RST, the notion of similarity relation and its use on a decision system, which provides when two objects are δ-similar and δ-discordant, with respect to a threshold δ.

Definition 3. *An* information system $(U, \mathcal{A})$ *is a tuple, where $U = \{x_1, \dots, x_n\}$ and $\mathcal{A} = \{a_1, \dots, a_m\}$ are finite, non-empty sets of objects and attributes, respectively. Each a in $\mathcal{A}$ corresponds to a mapping $\bar{a}\colon U \to V_a$, where V_a is the value set of a over U. For every subset B of $\mathcal{A}$, the B-indiscernibility relation*[1] *I_B is defined as the equivalence relation*

$$I_B = \{(x_i, x_j) \in U \times U \mid \text{ for all } a \in B,\ \bar{a}(x_i) = \bar{a}(x_j)\}, \tag{1}$$

where each class can be written as $[x]_B = \{x_i \mid (x, x_i) \in I_B\}$. I_B produces a partition on U denoted as $U/I_B = \{[x]_B \mid x \in U\}$.

In RST, data is represented as an information system. Given $A \subseteq U$, its lower and upper approximations w.r.t. B are defined by

$$I_B{\downarrow}A \;=\; \{x \in X \mid [x]_B \subseteq A\} \tag{2}$$

$$I_B{\uparrow}A \;=\; \{x \in X \mid [x]_B \cap A \neq \varnothing\} \tag{3}$$

[1] When $B = \{a\}$, i.e., B is a singleton, we will write I_a instead of $I_{\{a\}}$.

Definition 4. *A* decision system $(U, \mathcal{A} \cup \{d\})$ *is a special kind of information system, in which* d $\notin \mathcal{A}$ *is called the decision attribute, and its equivalence classes* $[x]_d$ *are called decision classes.*

A well-known approach to generate all reducts of a decision system is based on its discernibility matrix and function [8]. The discernibility matrix of $(U, \mathcal{A} \cup \{d\})$ is the $n \times n$ matrix O, defined by, for i and j in $\{1, ..., n\}$,

$$O_{ij} = \begin{cases} \varnothing & \text{if } d(x_i) = d(x_j) \\ \{a \in \mathcal{A} \mid \bar{a}(x_i) \neq \bar{a}(x_j)\} & \text{otherwise} \end{cases} \tag{4}$$

The *discernibility function* of $(U, \mathcal{A} \cup \{d\})$ is the map $f\colon \{0,1\}^m \to \{0,1\}$, defined by

$$f(a_1^*, ..., a_m^*) = \bigwedge \left\{ \bigvee O_{ij}^* \mid 1 \leq i < j \leq n \text{ and } O_{ij} \neq \varnothing \right\} \tag{5}$$

in which $O_{ij}^* = \{a^* \mid a \in O_{ij}\}$. The Boolean variables $a_1^*, \ldots, a_m^*$ correspond to the attributes from $\mathcal{A}$. It can be shown that the prime implicants of f constitute exactly all decision reducts of $(U, \mathcal{A} \cup \{d\})$.

We continue recalling the definition of similarity relationship, which extends the notion of equivalence relation and therefore the concept of equality.

Definition 5. *Given an arbitrary set* V*, the mapping* $E\colon V \times V \to [0,1]$*, is called* similarity relation *if it is reflexive, symmetric and transitive.*

In theory, we can define a similarity relation over the set of objects in an arbitrary way. However, in practice it is indeed resonable to refer to values of objects for available attributes.

There are several possibilities to define a similarity relation on the set of objects U. One of the most popular ways is as follows:

$$E_U(i,j) = \bigwedge_{a \in A} (E_a(a(i), a(j))) \tag{6}$$

Definition 6. *Given an information system* $\mathbb{A} = (U, \mathcal{A})$ *and a similarity relation family* $\mathcal{E} = \{E_a\colon V_a \times V_a \to [0,1] \mid a \in \mathcal{A}\}$ *we say that objects* $i, j \in U$ *are* ***δ-similar*** *if for all* $a \in \mathcal{A}$ *we have*

$$\delta \leq E_a(a(i), a(j))$$

with $\delta \in [0,1]$. *Otherwise, we say that objects* $i, j \in U$ *are* ***δ-discordant****, that is, if the following holds:* $\{a \in \mathcal{A} \mid E_a(a(i), a(j)) < \delta\} \neq \varnothing$.

3 Generalization of reducts and bireducts by similarities with Boolean decision attribute

In this section a threshold $\delta \in [0,1]$ is fixed, from which we will use the notions of δ-similar and δ-discordant pairs of objects to define the generalization of the discernibility function using similarity relations. Hence, an information system $\mathbb{A} = (U, \mathcal{A})$ and a similarity relation family $\mathcal{E} = \{E_a \colon V_a \times V_a \to [0,1] \mid a \in \mathcal{A}\}$ will also be fixed. Moreover, a linear ordering $\leq$ will also be fixed in U. Since the specific definition of the ordering is not important, any one can be considered. Given $i, j \in U$, we will say that $i < j$, if $i \leq j$ and they are not the same object.

First of all, the definitions of information reducts and bireducts are introduced.

Definition 7. *The set $B \subseteq \mathcal{A}$ is called* δ-information reduct *if and only if it is an irreducible subset such that every pair $i, j \in U$, which is δ-discordant by $\mathcal{A}$, is also δ-discordant by B.*

Definition 8. *The pair (B, X), where $B \subseteq \mathcal{A}$ and $X \subseteq U$, is called* δ-information bireduct *if and only if all pairs i, j of X are δ-discordant by B and the following properties hold:*

1. *There is no $C \subsetneq B$ such that all pairs $i, j \in X$ are δ-discordant by C.*
2. *There is no $X \subsetneq Y$ such that all pairs $i, j \in Y$ are δ-discordant by B.*

In this paper, the results will be focused on the general case of decision reducts and bireducts. The cases of δ-information reducts and bireducts arise as "particular cases" of them and similar results hold analogously. In this case we will have to make a distinction based on decision attribute because we have a definition whether the attribute is a Boolean decision or not. This section will handle decision systems with a Boolean decision attribute. i.e., Boolean decision systems.

Definition 9. *Let $\mathbb{A} = (U, \mathcal{A} \cup \{d\})$ be a Boolean decision system. The subset $B \subseteq A$ is called* δ-decision reduct *if and only if it is an irreducible subset such that all pair $i, j \in U$ is δ-discordant by B where $d(i) \neq d(j)$.*

Note that $d(i)$ and $d(j)$ are Boolean values. Next, the notion of decision bireduct is given.

Definition 10. *Let $\mathbb{A} = (U, \mathcal{A} \cup \{d\})$ be a Boolean decision system. The pair (B, X), where $B \subseteq \mathcal{A}$ and $X \subseteq U$, is called* δ-decision bireduct *if and only if every pair $i, j \in X$ is δ-discordant by B when $d(i) \neq d(j)$ and the following properties hold:*

1. *There is no $C \subsetneq B$ such that all pair $i, j \in X$ are δ-discordant by C, where $d(i) \neq d(j)$.*
2. *There is no $X \subsetneq Y$ such that all pair $i, j \in Y$ are δ-discordant by B, where $d(i) \neq d(j)$.*

Now, we are going to introduce the discernibility function in this general framework in order to obtain both δ-decision reducts and bireducts. Since for δ-decision reducts only the attributes are needed we will call it unidimensional δ-discernibility function (uni δ-d function) and for δ-decision bireducts, both attributes and objects are considered and so, we will call it bidimensional δ-discernibility function (bi δ-d function).

Definition 11. *Let* $\mathbb{A} = (U, A \cup \{d\})$ *be a decision system, the* unidimensional δ-discernibility function of $\mathbb{A}$, *is defined as the following conjunctive normal form (CNF):*

$$\tau_{\mathcal{A}}^{uni} = \bigwedge \Big\{ \bigvee \{a \in \mathcal{A} \mid E_a(a(i), a(j)) < \delta\} \mid i, j \in U,\ d(i) \neq d(j) \Big\}$$

where the elements of $\mathcal{A}$ *are the propositional symbols of the language. Also, we can denote it as:*

$$\tau_{\mathcal{A}}^{uni} = \bigwedge_{\{i,j \mid E_d(d(i),d(j))<\delta\}} \left(\bigvee_{\{a \mid d(i) \neq d(j)\}} a \right)$$

Note that, although the condition $i < j$ is not considered in the definition, this can be considered without loss of generality, since any proper closure is removed: If $i = j$, then $E_a(a(i), a(j)) = 1 \not< \delta$ and so, this case does not arises any clause. If $j < i$, then the same clause for $i < j$ is provided and so, this will be removed when the DNF will be computed. This remark can be applied to the rest of discernibility functions introduced in this paper.

Therefore, the unidimensional δ-discernibility function of $\mathbb{A}$ can be written as:

$$\tau_{\mathcal{A}}^{uni} = \bigwedge_{\{i,j \mid i<j, E_d(d(i),d(j))<\delta\}} \left(\bigvee_{\{a \mid d(i) \neq d(j)\}} a \right)$$

Next, the characterization of the δ-decision reducts is given.

Theorem 1. *Given a Boolean decision system* $\mathbb{A} = (U, A \cup \{d\})$. *An arbitrary set* B, *where* $B \subseteq \mathcal{A}$, *is a* δ*-decision reduct of* $\mathbb{A}$ *if and only if the cube* $\bigwedge_{b \in B} b$ *is a cube in the RDNF of* $\tau_{\mathcal{A}}^{uni}$.

The following definition is the natural extension of the discernibility function expression to δ-decision bireducts.

Definition 12. *Let* $\mathbb{A} = (U, A \cup \{d\})$ *be a decision system, the conjunctive normal form*

$$\tau_{\mathcal{A}}^{bi} = \bigwedge \{ i \vee j \bigvee \{a \in \mathcal{A} \mid E_a(a(i), a(j)) < \delta\} \mid i, j \in U, i < j,\ d(i) \neq d(j)\}$$

where the elements of U *and* $\mathcal{A}$ *are the propositional symbols of the language, is called the* bidimensional δ-discernibility function.

The following theorem characterize the δ-decision bireducts.

Theorem 2. *Given a decision system* $\mathbb{A} = (U, A \cup \{d\})$, *an arbitrary pair* (B, X), $B \subseteq \mathcal{A}$, $X \subseteq U$, *is a* δ*-decision bireduct if and only if the cube* $\bigwedge_{b \in B} b \wedge \bigwedge_{i \notin X} i$ *is a cube in the RDNF of* $\tau_{\mathbb{A}}^{bir}$.

4 Conclusion and future work

We have studied the reducts and bireducts in the classic environment of RST and considering similarity relations. We have generalized discernibility function, from which we could get the reducts and bireducts in these environments.

The inclusion of the similarity relations in theory provides a greater flexibility in these environments, dramatically increasing the range of possible applications. Moreover, we have also considered the δ-information reducts and bireducts in FCA, providing a new reduction method based on RST, which very close to the FCA framework.

As future work, we will extend the theory to obtain bireducts in fuzzy environments, such as in fuzzy rough sets [1, 2]. Moreover, we will study in depth in the relation between concept lattice reduction and rough set reduction considering similarity relations and in the general fuzzy case. Furthermore, we apply the theory developed in both theories to practical cases.

References

1. C. Cornelis, R. Jensen, G. Hurtado, and D. Ślęzak. Attribute selection with fuzzy decision reducts. *Information Sciences*, 180:209–224, 2010.
2. C. Cornelis, J. Medina, and N. Verbiest. Multi-adjoint fuzzy rough sets: Definition, properties and attribute selection. *International Journal of Approximate Reasoning*, 55:412–426, 2014.
3. L. Fariñas del Cerro and H. Prade. Rough sets, twofold fuzzy sets and modal logic—fuzziness in indiscernibility and partial information. In A. D. Nola and A. Ventre, editors, *The Mathematics of Fuzzy Systems*, pages 103–120. Verlag TUV Rheinland, 1986.
4. A. Janusz and D. Ślęzak. Rough set methods for attribute clustering and selection. *Applied Artificial Intelligence*, 28(3):220–242, 2014.
5. N. Mac Parthalain and R. Jensen. Simultaneous feature and instance selection using fuzzy-rough bireducts. In *Fuzzy Systems (FUZZ), 2013 IEEE International Conference on*, pages 1–8, July 2013.
6. J. Medina. Relating attribute reduction in formal, object-oriented and property-oriented concept lattices. *Computers & Mathematics with Applications*, 64(6):1992–2002, 2012.
7. Z. Pawlak. Rough sets. *International Journal of Computer and Information Science*, 11:341–356, 1982.
8. A. Skowron and C. Rauszer. The discernibility matrices and functions in information systems. In R. Słowiński, editor, *Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory*, pages 331–362. Kluwer Academic Publishers, 1992.
9. D. Ślęzak and A. Janusz. Ensembles of bireducts: Towards robust classification and simple representation. In T.-h. Kim, H. Adeli, D. Ślęzak, F. Sandnes, X. Song, K.-i. Chung, and K. Arnett, editors, *Future Generation Information Technology*, volume 7105 of *Lecture Notes in Computer Science*, pages 64–77. Springer Berlin Heidelberg, 2011.
10. S. Stawicki and D. Ślęzak. Recent advances in decision bireducts: Complexity, heuristics and streams. *Lecture Notes in Computer Science*, 8171:200–212, 2013.

Reasoning on Molecular Interaction Maps

Jean-Marc Alliot, Robert Demolombe, Luis Farinas, Martín Diéguez and Naji Obeid

University of Toulouse, CNRS, IRIT
Toulouse, France

Abstract. Metabolic networks, formed by a series of metabolic pathways, are made of intra-cellular and extracellular reactions that determine the biochemical properties of a cell, and by a set of interactions that guide and regulate the activity of these reactions. Cancer, for example, can sometimes appear in a cell as a result of some pathology in a metabolic pathway. Most of these pathways are formed by an intricate and complex network of chain reactions, and are often represented in *Molecular Interaction Maps* (MIM), a graphical, human readable form of the cell cycle checkpoint pathways. In this paper, we present a logic, called Molecular Interaction Logic, which semantically characterizes MIMs.

1 Introduction

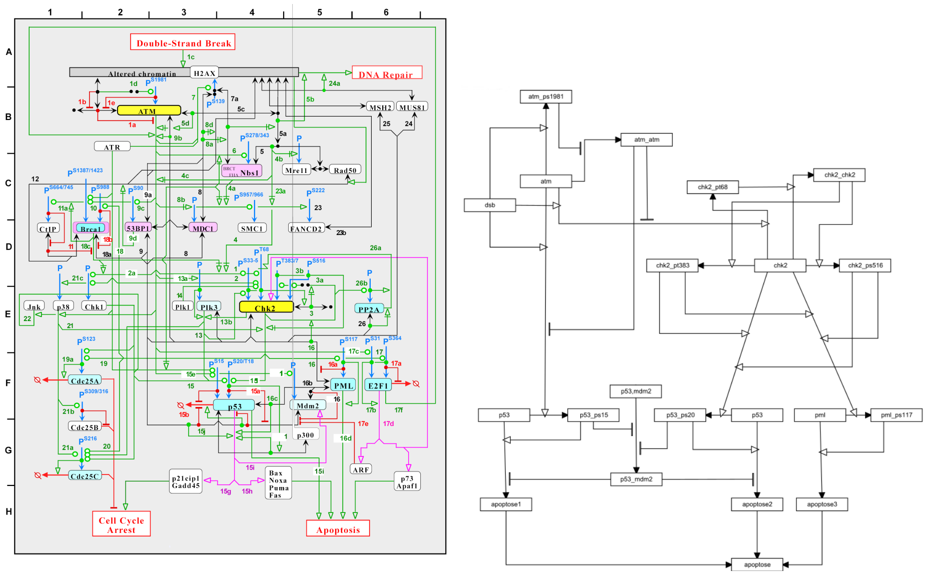
Metabolic networks, formed by series of metabolic pathways, are made of intra-cellular and extracellular reactions that determine the biochemical properties of a cell, and by a set of interactions that guide and regulate the activity of these reactions. These reactions can be positive (production of a new protein) or negative (inhibition of a protein in the cell). These reactions are at the center of a cell's existence, and are modulated by other proteins, which can either enable these reactions or, on the opposite, inhibit them.

Medical and pharmaceutical researches [11, 8] showed that the break of the double strand of DNA sometimes appear in a cell as a result of some pathology in a metabolic pathway, and double strand break (*dsb*) is a major cause of cancer.

These pathways are used to investigate the molecular determinants of tumor response in cancers. The molecular parameters include the cell cycle checkpoint, DNA repair and apoptosis[1] pathways [15, 11, 8, 12, 14]. When DNA damage occurs, cell cycle checkpoints are activated and can rapidly kill the cell by apoptosis or arrest the cell cycle progression to allow DNA repair before cellular reproduction or division. Two important checkpoints that appear to function when parallel transduction cascades from DNA damage to the cell cycle checkpoint effectors are the atm-$chk2$ and the atr-$chk2$ pathways [15].

Most of these pathways are formed by an intricate and complex network of chain reactions, and are often represented in *Molecular Interaction Maps* (MIM), a human readable form of the cell cycle checkpoint pathways, such as the one in Figure 1(a), which represents the atm-$chk2$ and atr-$chk2$ pathways cited above.

[1] Apoptosis is the process of programmed cell death.



(a) atm-$chk2$/atr-$chk1$ molecular interaction map.

(b) $atm - chk2$ pathway

MIMs become increasingly larger and their density is constantly enriched with new information (references, date, authors, etc.). Although essential for knowledge capitalization and formalization, MIMs are difficult to use because of the very large number of elements involved as well as the inherent knowledge which, sometimes, is not formally described in the map.

In this paper we present a method to transform a MIM into a set of logical formulas. Subsets of Figure 1(a) will be used as examples, concentrating on the modelling of the atm-$chk2$ pathway leading to apoptosis.

The rest of this paper is organized as follows: section 2 introduces the concept of Molecular Interaction Maps and how they can be translated into a set of logical formulas. Section 3 describes Molecular Interaction Logic, a logic which is capable of describing and reasoning about general pathways and finally section 4 ends the paper with conclusions and future work.

2 Molecular Interaction Maps

A Molecular Interaction Map [10] (MIM) is a diagram convention which represents the interaction networks of multi-protein complexes, protein modifications and enzymes that are substrates of other enzymes. Although interactions between elements of a MIM can be complex, they can be represented using only three basic connectors: production

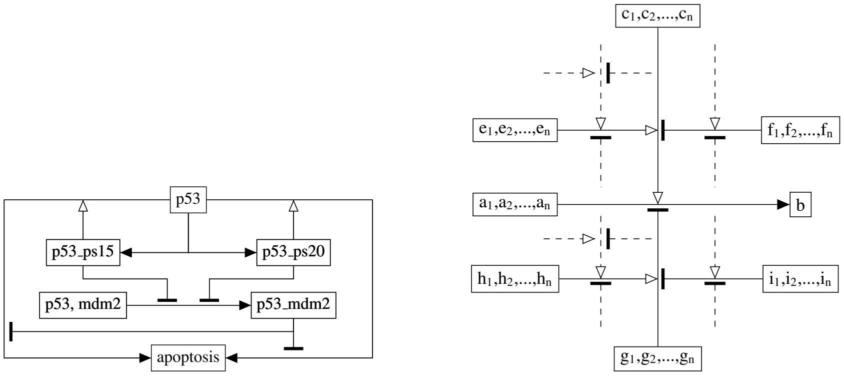
($\rightarrow$), activation ($\rightarrow\!\triangleright$) and inhibition ($\dashv$). Figure 1(b) presents the atm-$chk2$ pathway, an already pretty complex part of the large MIM of Figure 1(a), using only the aforementioned connectors.

A production relation means that a new substance is created as a result of a reaction on several primary components. For instance, the protein atm can be dimerized to become the atm_atm protein or phosphorylated at serine 1981 resulting in the production of atm_ps1981. These reactions can be triggered or blocked by other proteins or conditions. For example, in Figure 1(b), atm_ps1981 blocks the dimerization of atm into atm_atm, while the double strand break (dsb) of DNA triggers the production of atm_ps1981 by atm.

These interactions can be "stacked": for example, protein $p53$ can be phosphorylated at serine 15 to become $p53\_ps15$ (see Figure 1(b)). This reaction is triggered by atm, but the triggering itself has to be activated by dsb and can be blocked by atm_atm. Thus, the two main actions (production of a protein or inhibition of a protein) can be triggered or blocked by a stack of preconditions.

2.1 Translation of MIMs into formulas

Our first goal is to translate any MIM into a set of logical expressions in order to perform several automated reasoning tasks such as deduction or abduction. First, focusing on the diagram of Figure 1(c) (which corresponds to a sub-diagram of Figure 1(b)) will help getting an intuitive idea of how translation is performed.



(c) Apoptosis by $p53\_ps20$ and $p53\_ps15$ mediation.

(d) The general form of a basic production.

Here $apoptosis$ arises when protein $p53$ is phosphorylated at serine 20 or 15 (instances $p53\_ps20$ and $p53\_ps20$ respectively). However, $apoptosis$ would not happen if the dimer $p53\_mdm2$ is present. Thus the context would be *if* $p53$ *and either* $p53\_ps20$

or $p53\_ps15$ *are present and* $p53\_mdm2$ *is absent then* $apoptosis$ *is produced* (this example should of course be completed with the rules for producing the rest of objects in the diagram).

The general form of production relations is displayed in Figure 1(d).

Each arrow can be either an activation or an inhibition of the relation it applies to, and these activations/inhibitions can be stacked on any number of levels. The above examples give the idea behind the translation: it is a recursive process starting from the production relation and climbing up the tree. In order to formally describe the translation, the concept of *pathway context* is now defined:

Definition 1 (Pathway context). *Given a set of entities, a pathway context is formed by expressions defined by the following grammar:*

$$\alpha ::= \langle \alpha P \rightarrowtriangle, \alpha Q \dashv \rangle | \langle P \rightarrowtriangle, Q \dashv \rangle,$$

where P and Q are sets (possibly empty) of propositional variables representing the conditions of activation ($\rightarrowtriangle$) or inhibition ($\dashv$) of the reaction. The first part of the pair is the activation context, the second part is the inhibition context. One, or both sets can be empty. ■

For example, the $p53 \rightarrow apoptosis$ reaction of Figure 1(c) would lead to the following two pathway contexts:

$$\langle p53\_ps20 \rightarrowtriangle, p53\_mdm2 \dashv \rangle \tag{1}$$

$$\langle p53\_ps15 \rightarrowtriangle, p53\_mdm2 \dashv \rangle \tag{2}$$

Definition 2 (Activation and inhibition expressions). *Given a pathway context $\alpha = \langle \alpha' P \rightarrowtriangle, \beta' Q \dashv \rangle$, the* activation *and the* inhibition expressions *associated with the context α (denoted by $A(\alpha)$ and $I(\alpha)$) are defined recursively as:*

$$A(\alpha) = \bigwedge_{p \in P} p \wedge A(\alpha') \wedge (\bigvee_{q \in Q} \neg q \vee I(\beta')) \qquad I(\alpha) = \bigvee_{p \in P} \neg p \vee I(\alpha') \vee (\bigwedge_{q \in Q} q \wedge A(\beta'))$$

The above expressions define the general forms of $A(\alpha)$ and $I(\alpha)$. If one part of the context α is empty, then the corresponding part is of course absent in $A(\alpha)$ and $I(\alpha)$. ■

Following such definition, formulas associated with (1) are:

$$A((1)) = p53\_ps20 \wedge \neg p53\_mdm2 \qquad I((1)) = \neg p53\_ps20 \vee p53\_mdm2$$

Definition 3 (Causal pathway formulas). *A* causal pathway formula *is defined by the following grammar:*

$$F ::= [\alpha](p_1 \wedge \cdots \wedge p_n \rightarrow \mathbf{Pr}\, q) \mid [\alpha](p_1 \wedge \cdots \wedge p_n \rightarrow \mathbf{In}\, q) \mid F \wedge F,$$

where α *is a pathway context,* $p_1, \cdots, p_n$*,* q *are propositional variables while* $\mathbf{Pr}$ *and* $\mathbf{In}$ *are modal concepts that qualify the process of activation or inhibition of proteins.* ■

Applied to the example of Figure 1(c), the causal pathway formula associated with the production rule $p53 \twoheadrightarrow apoptosis$ is

$$[(1)](p53 \rightarrow \mathbf{Pr}\ apoptosis) \wedge [(2)](p53 \rightarrow \mathbf{Pr}\ apoptosis)\,. \tag{3}$$

Observation 1 *Each MIM can now be represented in terms of a causal pathway formula.* ■

3 Molecular Interaction Logic

In this section the semantics of the Molecular Interaction Logic (MIL) is formally introduced. This work extends a previous one [5, 4] where the MIMs were formalized via first order logic with equality, in which the pathway contexts were limited to one level of depth. From now on, p means *protein p is present* and $\neg p$ means *protein p is absent*.

Definition 4 (MIL interpretation). *A MIL interpretation consists of a pair* (V_1, V_2) *of classical evaluations i.e.* $V : \mathcal{P} \rightarrow \{True, False\}$ *where* $\mathcal{P}$ *is the set of propositional variables.* ■

The intuitive meaning behind these two evaluations correspond for V_1 to the protein present or absent, and for V_2 to the state of the protein resulting from the chemical reactions in the cell[2].

Definition 5 (Satisfaction relation). *Given a MIL interpretation* (V_1, V_2) *and a formula* α*, the satisfaction relation is defined as:*

1) $(V_1, V_2) \vDash p$ *iff* $V_1(p) = True$ *for* $p \in \mathcal{P}$

2) $\wedge$, $\vee$ *and* $\rightarrow$ *are satisfied as in classical logic.*

3) $(V_1, V_2) \vDash \mathbf{Pr}\, p$ *iff* $V_1(p) = V_2(p) = True$

4) $(V_1, V_2) \vDash \mathbf{In}\, p$ *iff* $V_1(p) = V_2(p) = False$

[2] If the semantics of the modal logic S5 is restricted to have at most two worlds then a strong normal form in which conjunctions and disjunctions are not in the scope of a modal operator can be found for this new logic [1]: the pathway causal formulas of MIL verify this condition.

5) $(V_1, V_2) \models [\alpha]F$ *iff* $(V_1, V_2) \not\models A(\alpha)$ *or* $(V_1, V_2) \models F$

■

As usual, a formula F is satisfiable if there is a model (V_1, V_2) such that $(V_1, V_2) \models F$.

Observation 2 *MIL can be characterized by the axioms of classical logic, plus the axioms:*

1. $[\alpha]F \leftrightarrow (A(\alpha) \rightarrow F)$
2. $\mathbf{Pr}\ p \rightarrow p$*, if* p *is produced then* p *is present*
3. $\mathbf{In}\ p \rightarrow \neg p$*, if* p *is inhibited then* p *is absent*

■

As a result of MIL semantics, the causal pathway formula (3) is logically equivalent to the conjunction of the following implications:

$$(p53 \wedge p53\_20 \wedge \neg p53\_mdm2) \rightarrow \mathbf{Pr}\ apoptosis \quad (4)$$

$$(p53 \wedge p53\_20 \wedge \neg p53\_mdm2) \rightarrow \mathbf{Pr}\ apoptosis \quad (5)$$

Observation 3 *Any MIM can be transformed into a causal pathway formula, and every causal pathway formula is equivalent to a boolean composition of:*

- *propositional variables or their negation*
- *propositional variables qualified by* $\mathbf{Pr}$ *or* $\mathbf{In}$ *or their negation*

■

Axioms 2) and 3) of observation 2 have as consequence:

Observation 4 *Given a MIM formula* F*, adding* $\mathbf{Pr}\ p \rightarrow p$ *and* $\mathbf{In}\ p \rightarrow \neg p$ *for each propositional variable* p *in* F*, enables us to embbed MIL into classical logic.* ■

The notions of completion and production axioms, which are both important and implicit in MIMs, are presented first.

4 Conclusions and future work

We have presented a method to automatically translate MIMs into logical formulas, formalism that allows performing several kinds of reasoning such as deduction (in order to find inconsistencies in a representation) and abduction (which allows answering queries asked on MIMs). As a future work we want, on one hand, to enrich the language of MIL, with concepts like ”aboutness” which are able to qualify, for example, proteins, allowing us to isolate the subgraph of a given MIM, regarding the qualified proteins. On the other hand, such enrichment of the language could include the introduction of temporal operators while to incorporate a temporal aspect to MIMs

References

1. del Cerro, L.F., Herzig, A.: Contingency-based equilibrium logic. In: LPNMR'11. pp. 223–228 (2011)
2. Demolombe, R., Farinas, L.: An Inference Rule for Hypothesis Generation. In: IJCAI'91 (1991)
3. Demolombe, R., Farinas, L.: Information about a given entity: From semantics towards automated deduction. Journal of Logic and Computation 20(6), 1231–1250 (2010)
4. Demolombe, R., Farinas, L., Obeid, N.: Logical model for molecular interactions maps. In: Logical Modeling of Biological Systems, pp. 93–123. John Wiley & Sons (2014)
5. Demolombe, R., L. Fariñas, Obeid, N.: Translation of First Order Formulas into Ground Formulas via a Completion Theory. Journal of Applied Logic (), to appear
6. Erwig, M., Walkingshaw, E.: Causal reasoning with neuron diagrams. In: VLHCC '10. pp. 101–108 (2010)
7. Farinas, L., Inoue, K. (eds.): Logical Modeling of Biological Systems. John Wiley & Sons (2014)
8. Glorian, V., Maillot, G., Poles, S., Iacovoni, J.S., Favre, G., Vagner, S.: Hur-dependent loading of mirna risc to the mrna encoding the ras-related small gtpase rhob controls its translation during uv-induced apoptosis. Cell Death & Differentiation 18(11), 1692–70 (2011)
9. Inoue, K.: Linear resolution for consequence finding. Artificial Intelligence 56(2-3), 301 – 353 (1992)
10. Kohn, K.W., Aladjem, M.I., Weinstein, J.N., Pommier, Y.: Molecular interaction maps of bioregulatory networks: A general rubric for systems biology. Molecular Biology of the Cell 17(1), 1–13 (2006)
11. Kohn, K.W., Pommier, Y.: Molecular interaction map of the p53 and mdm2 logic elements, which control the off-on swith of p53 response to dna damage. Biochemical and biophysical research communications 331(3), 816–27 (2005)
12. Lee, W., Kim, D., Lee, M., Choi, K.: Identification of proteins interacting with the catalytic subunit of pp2a by proteomics. Proteomics 7(2), 206–214 (2007)
13. Muggleton, S., Bryant, C.H.: Theory completion using inverse entailment. In: ILP'00. pp. 130–146 (2000)
14. Pei, H., Zhang, L., Luo, K., Qin, Y., Chesi, M., Fei, F., Bergsagel, P.L., Wang, L., You, Z., Lou, Z.: MMSET regulates histone H4K20 methylation and 53BP1 accumulation at DNA damage sites. Nature 470(7332), 124–128 (2011)
15. Pommier, Y., Sordet, O., Rao, V.A., Zhang, H., Kohn, K.W.: Targeting chk2 kinase: molecular interaction maps and therapeutic rationale. Current pharmaceutical design 11(22), 2855–72 (2005)
16. Ray, O., Whelan, K., King, R.: Logic-based steady-state analysis and revision of metabolic networks with inhibition. In: CISIS'10. pp. 661–666 (2010)
17. Reiser, P.G., King, R.D., Kell, D.B., Muggleton, S., Bryant, C.H., Oliver, S.G.: Developing a logical model of yeast metabolism. Electronic Transactions in Artificial Intelligence 5, 233–244 (2001)
18. Rougny, A., Froidevaux, C., Yamamoto, Y., Inoue, K.: Analyzing SBGN-AF Networks Using Normal Logic Programs. In: Logical Modeling of Biological Systems, pp. 44–55. John Wiley & Sons (2014)