Jornada de Seguimiento de Proyectos, 2005
Programa Nacional de Tecnologías Informáticas

The incorporation of the Mathematical Morphology, the Formal Concept Analysis and the Fuzzy Logic techniques and tools to the data cleaning, aggregating, reducing and mining stages of the Knowledge Discovery process TIC2003-08693

Ramón Fuentes-González*
Dpto. de Automática y Computación
Universidad Pública de Navarra
31006-Pamplona

Abstract

The principal aim of this project is to set up new mathematical models together with efficient associated algorithms that allow the incorporation of some elements of the Formal Concept Analysis (FCA), of the Mathematical Morphology (MM), of the Aggregation Functions Theory (AFT) and of the Fuzzy Relation Theory (FRT) in the following phases included in the Knowledge Discovery (KD) processes: data cleaning, data aggregating, data reducing and data mining (DM). Also, this project is intended to compare our algorithms with others associated with the classical methods in DM.

Keywords: Knowledge Discovery, Data Mining, Fuzzy Concept Analysis, Formal Concept Analysis, Fuzzy Mathematical Morphology, Fuzzy Relations.

1 The aim of the project

The Knowledge Discovery in databases (KDD) [1] process consists of the automated extraction of patterns representing knowledge implicitly stored in large databases, date warehouses and other massive information repositories. It is considered that the KDD process consists of an iterative sequence of the following steps: data cleaning, data integration (data grouping and data reduction), data selection, data transformation, data mining, pattern evaluation and knowledge representation [2].

* Email: rfuentes@unavarra.es

At present time, the KDD process is a multidisciplinary field [3], drawing work from areas including database technology, artificial intelligence, machine learning, neural networks, statistics, optimization, pattern recognition, fuzzy logic, soft computing, formal concept analysis [3], rough sets theory, knowledge-based systems, knowledge acquisition, information retrieval, high-performance computing and data visualization.

With this project it is tried to fulfil, in collaboration with the aforementioned ones, two new incorporations to that discipline list: the Mathematical Morphology (Crisp Theory [5] and Fuzzy Theory [6], [7], [8]) and the Fuzzy Concept Analysis [9], [10]. Specifically, our main objective is to set up some mathematical models and its associate algorithms that allow to incorporate elements of the disciplines: Fuzzy Concept Analysis, Fuzzy and Crisp Mathematical Morphology, Aggregation Functions Theory [11], and Fuzzy Relations Theory [13], in the following KDD steps: data cleaning, data grouping, data reduction and data mining.

The attainment of this objective will suppose the incorporation of some new ideas and methods in processes of KDD and Data Mining, as well as their comparison with the present mechanisms of KDD that are developed from the aforementioned different disciplines.

2 Level of reached success in the project

Next, a summary (including the fulfilment of the proposed tasks and the developed activities), is shown.

1. Fulfilment of the initially proposed plan of work.

The scheme of work associated to the used methodology which we proposed in the Scientist-technique Memory of the Project is specified in 14 tasks. To date, the corresponding part of that scheme includes the first eleven tasks. Next, we expose a summary of the made activities, detached by tasks.

Task 01: *Compilation of specialized information that concerns the project* and Task 02: *Analysis and evaluation of generic applications of* DATA MINING (DM). (Whose development was programmed during the first year of the project, 2004).

In the main, the work of the group of investigation related to these two tasks has been completed within the expected term, and it has consisted of:

01a) Compilation of the basic bibliography and of the works of investigation published in KNOWLEDGE DISCOVERY IN DATABASES (KDD) and DATA MINING (DM) specialized journals, fundamentally the one that more directly appears related to the topics including in the memory of the project. Although this task is considered fundamentally complete, it is open to the foreseeable updates that we will do during the period that remains to finalize the project.

The listing of those resources will be included in the final memory, specifying those that have influenced, inspired and complemented the results and investigations related to this project.

01b) Web location of permanent resources over subjects, investigations, events, new bibliography, software and groups of investigation related to the topics: KDD and DM in general, as well as in the specific case of resources in the field of THE FORMAL CONCEPT ANALYSIS applied to the extraction of knowledge in databases.

One of the main achievements obtained as a result of this task has been the incorporation of our group to the Spanish Network of Data Mining. This organization incorporates the most of the DM and KDD working groups in the Spanish Public Universities.

In addition, this activity has allowed to the group to get up to the lists of interest related to publications, events, meetings and location of new resources.

In the final memory of this project, a listing of these resources will be included, along with the details and explanations that we will consider pertinent.

02). With regard to the DM and KDD specific tools analysis, the following software was selected:

Utilities of data analysis with possibilities of graphical representation of the data or summaries of them:

- *3DV8 Enterprise Edition, Open VisiCube, Miner3D ACCESS, S-Plus* (language), *Axum 7.*

The three first utilities implement standards of the graphical data mining, so they will serve us as models for the algorithms that we will develop in the following phases of the project.

So much the *S-Plus* as *Axum 7* utilities have the possibility of implementing algorithms that directly manipulate data tables. (The frequent support of the data presentation in DM techniques).

All these applications have analyzed in Windows, although analogous utilities to S-Plus are available in UNIX and LINUX.

(i) *S*pecifically designed utilities for data mining tasks:

- *Clementine, Insightful Miner 3.0, Weka-3-2.*

The *Clementine* and *Insightful Miner 3,0* utilities are selected since they implement the majority of the classic data mining algorithms.

We have selected those utilities in order to compare our algorithms, with the classic ones, these last implemented in *Clementine* and *Insightful Miner.*

Both are similar utilities, but the utility *Insightful Miner 3,0* has the advantage to use the S-Plus language directly.

Finally, *the Weka-3-2* is a specific tool for the investigation in DM. It is a free access utility developed in Java. We have selected it since it can become a standard of the algorithms on KDD and DM, at least at academic level, both in investigation and in implementation.

(ii) Specific languages for the data tables manipulation:

- *S-Plus* (language), *MICROSOFT ACCESS* (version 2003).

Since it has been mentioned in the section (i), we think that in addition to the graphical possibilities for the representation of data, the *S-Plus* language seems suitable for the implementation of algorithms that transform, (via morphologic operators), a date table in other one.

The initial files (on which the developed algorithms that implement the morphologic operators have been proven), have been recovered by means of the utility *Microsoft Access*, so the first versions of our algorithms are implemented in Access. Next, the S-Plus language will be used to obtain the later versions, since the structure of this last language will make a more efficient version of those algorithms.

(iii) Complementary utilities:

- *Toscanaj , Fuzzy Query, GUHA.*

The Toscanaj utility is a new version of *Toscana*, the standard utility for the analysis of data associated to contexts using the Theory of *Formal Concept Analysis* methods *(FCA)* This new *Toscanaj* version is implemented in Java, reason by which its integration with *Weka* is a feasible task that we will try to approach when the tables of instances will be interpreted as contexts.

We consider that the utility *Fuzzy Query* is useful to introduce an aspect that non appears in the specialized literature, and that we want to implement in our algorithms: the consideration of the *Fuzzy Logic* (*FL*) as a basic element for the models containing vagueness, circumstance that appears frequently in database analysis. This utility is designed specifically for fuzzy queries in structured databases (FSQL).

Finally, *GUHA* is a free access utility with implemented algorithms of analysis of conceptual data, algorithms that are of our interest in this project.

Let us indicate here that this process of search of suitable tools for our investigation, although sufficiently defined, it is not closed. We have verified in our access to the specialized sources on that the increasing interest in KDD subjects has done that many investigators of very diverse matters try to incorporate they tools to problems of data mining.

Task 03: *Design of the mathematical model for the morphologic filters in data mining* and Task 04: *Expression of the morphologic filters for the phases of data cleaning and data grouping in terms of the Mathematical Morphology and the theory of Fuzzy Relations.* (Whose development was programmed during the year 2004 and the beginning of the 2005).

This is the task in which we are working now. In particular, we have designed the algorithms that implement the basic filters for the treatment of data tables. Such basic filters are the *erosions* and *dilatations* of data tables, (represented by fuzzy subsets or fuzzy relations), using another fuzzy or crisp relation as a *neighbourhood relation* (or *structuring relation*).

Later, and following the techniques of the *Fuzzy Mathematical Morphology Theory* used in another context (the one of treatment of images [14]), the combination of those basic filters they have provided new filters such as *openings*, *closings*, *hit or miss operators*, etc.

Summary of results obtained in the development of the activities of tasks 03 and 04

With regard to tasks 03 and 04, we have developed and extended the ideas that previously appeared outlined in a published work [12] by members of this group. The purpose of this work is the application of the Mathematical Morphology tools and methods to handle the useful and possibly hidden information of the fuzzy and crisp relational systems.

In order to introduce the basic filters erosion and dilatation of data tables, (fuzzy sets $A \in L^E$, $B \in L^E$,… or fuzzy relations $S \in L^{E\times F}$), by a neighbourhood relation, (a fuzzy relation $R \in L^{E\times E}$), we consider the usual composition operators $\circ_\top$ and $\lhd_\rightsquigarrow$ in Fuzzy Logic [9] defined by:

$$(R \circ_\top B)(x) = \sup_{y\in F} \, (R(x,y) \top B(y)),$$

$$(A \circ_\top R)(y) = \sup_{x\in E} \, (A(x) \top R(x,y)).$$

$$(R \lhd_\rightsquigarrow B)(x) = \inf_{y\in F} \, (R(x,y) \rightsquigarrow B(y)),$$

$$(A \lhd_\rightsquigarrow R)(y) = \inf_{y\in F} \, (A(x) \rightsquigarrow R(x,y)).$$

$$(R \circ_\top S)(x,z) = \sup_{y\in F} \, (R(x,y) \top S(y,z)),$$

$$(R \lhd_\rightsquigarrow S)(x,z) = \inf_{y\in F} \, (R(x,y) \rightsquigarrow S(y,z)).$$

Where L is:

(i) A complete chain in *[0,1]* with the usual order or

(ii) a complete lattice of closed intervals *[a,b]* of *[0,1]* with de usual induced order.

And where $\top$ and $\rightsquigarrow$ represent a t-norm and a fuzzy implication in L respectively [10].

With those elements, we extend the classical crisp Minkowski operators [11] in $\mathbb{R}^2$: sum (⊕), and difference (⊖) used in Mathematical Morphology [14], to fuzzy subsets by

$$A \oplus R = R \circ_\top A \quad \text{or} \quad S \oplus R = R \circ_\top S \quad \text{and}$$

$$A \ominus R = R \lhd_\rightsquigarrow A \quad \text{or} \quad S \ominus R = R \lhd_\rightsquigarrow S.$$

And finally, we introduce the basic fuzzy morphological filters for data mining:

Erosion of the fuzzy set A by the structuring relation R : $\mathcal{E}(A, R) = A \ominus R^{op}$, been R^{op} the opposite relation : $R^{op}(x,y) = R(y,x)$

Dilatation of the fuzzy set A by the structuring relation R : $\mathcal{D}(A, R) = A \oplus R$.

Opening and *closing* filters as composition of the last ones:

Opening of the fuzzy set A by the structuring relation R : $@(A, R) = \mathcal{D}(\mathcal{E}(A, R), R)$.

Closing of the fuzzy set A by the structuring relation R : $C(A, R) = \mathcal{E} (\mathcal{D} (A, R), R)$.

We have shown some interesting result about this operators in the data mining context.

Let us indicate finally in this section that we are introducing new operators (*difference operators)* as differences of the basic ones. Later we will undertake the study of new definitions that implement in our context the concepts of *skeleton* and *skeleton by zones of influence* [11], that are used in mathematical morphology in grouping processes.

With this idea , and since the skeleton concept is associate to *distances* between items, our present task is to analyze bibliography of distances defined between fuzzy and crisp subsets of instances.

Also we want to make notice that we analyze data tables in information systems, we have observed certain similarity between the topics of our investigations and the techniques of the Theory of "*Rough Sets*" (RS) of *Pawlak* [15]. This is a task that we did not contemplate in our initial project. But our intention is to make a meticulous analysis of those similarities, with the purpose of seeing the possibility of including the results of RS in our investigation.

Task 05: *Design of algorithms associated to the morphologic filters for data cleaning, data grouping and data mining* and Task 06: *First implementation of the algorithms, proof and debugger them* (planned between June of 2004 and July of 2005).

We have developed the algorithms for the formal expressions of the following filters: *erosion*, *dilatation*, *opening*, *closing*, *hit or miss* and *difference operators*. At date, this implementation has been made for crisp and fuzzy subsets and with crisp relations as structuring elements only . Those algorithms are made as generalized products of matrices. Later, we will implement the most complex case of filters obtained with structuring fuzzy relations that are not crisp, problem that it should be undertaken after making an analysis on the more favourable type of implication and t-norm operators. This one seems to be a laborious task, although it seems that, like in other many fields related to the Fuzzy Logic, the operators of *Lukasiewicz* are the most appropriate to define the erosions and dilatations basic filters.

Taking a free access database as an element of test with near 50,000 instances and 14 attributes, (database *Adult* or *Census Income* with information on the census of the United States in 1994), and using Microsoft Access, we have made one first implementation of a great part of the mentioned filters. At the moment, we are dedicated to the interpretation of these results.

At the same time, we are implementing the algorithms to those filters in the mentioned language S-PLUS.

These tasks will continue until end of October of 2005.

Task 07: *Expression of the procedures of data grouping and data mining in terms of the Fuzzy Concepts Analysis and the theory of Fuzzy Relations*, Task 08: *Design of algorithms associated to the previous procedures* and Task 09: *First implementation of the previous algorithms, proof and debugger of them.* (planned between January of 2005 andl October of 2005).

In this section we have obtained some results that allow us to extract information from a fuzzy context in which there are unknown values. In this line, and in collaboration with the *Dpto of Matemática Aplicada de la Universidad del Pais Vasco*, we have proposed a method (using the idea of frequent sets) for the elimination of objects (instances) or attributes (fields) in fuzzy contexts in which these absent values appear. Moreover, we will set up the implications between intervals to replace the absent values and to limit the lack of information in the context.

Finally, we have used the idea of frequent items to define the frequent L-Fuzzy concepts that give us a method to organize the information obtained from a fuzzy context and we have also used the association rules to set up implications between the attributes of the fuzzy context.

As a result of these investigations we have presented a paper in the congress Estylf 04 in September of 2004 and other two have been acepted, one of them in the congress: *EUSFLAT-LFA 2005 Joint Conference* and another in *TAMIDA 2005 (III Taller Nacional de Minería de Datos y Aprendizaje*, organized by the Spanish network of Data Mining and Learning (CICYT Tic2002-11124-e). Both will be celebrated in September of this year 2005. The references of these works appear in the following section.

Task 10: *Study of the process of the knowledge fusion using Aggregation's Functions* and Task 11: *Design of algorithms associated to the previous procedures.* (Whose development was planned in the period April of 2005 until December of 2005).

The development of these activities will become in October of 2005, since the previous ones have lasted more than planned.

3 Indicators of the results

To date, and like result of the developed work related with our investigations, we have published the following communications:

1. A. Burusco, R. Fuentes-González and C. Alcalde. Use of rules of association in the theory of L-Fuzzy Concepts. *Estylf* 2004. Jaén (2004)

2. Alcalde, A. Burusco and R. Fuentes-González. Treatment of contexts with absent values. (admitted in TAMIDA 2005, III national Factory of Mining of Data and Learning, organized by the Red Española de Minería de Datos y Aprendizaje , Granada. September of 2005).

3. C. Alcalde, A Burusco and R. Fuentes-González. Treatment of the incomplete information in L-fuzzy contexts. (admitted in Eusflat-Lfa 2005 joint Conference, Barcelona. September 2005).

On the other hand, in a term of two months we will be able to send to the specialized journals, two works with the results that we have obtained related to the morphologic operators applied to data tables. These works will turn on the implementation of the filters mentioned in the summary that there are including in this report of the results corresponding to tasks 03 and 04.

4 References

[1] Frawley, W,J., Piatetsky-Shapiro, G. and Matheus, C.J. Knowledge discovery databases: An overview. In: Piatetsky-Shapiro, G. and Frawley, W,J.(eds) Knowledge Discovery in Databases, AAAI/MIT, (1991) 1-27

[2] Han J. and Kamber M. Data Mining. Concepts and Thecniques. Academic Press 2001

[3] Hand, D., Mannila, H., Smyth, P. Principles of Data Mining. The MIT Press. 2001

[4] Ganter, B., Wille, R. Formal Concept Analysis. Mathematical Foundations. Springer 1999.

[5] Serra J. Mathematical Morphology Vol 1 and 2 Academic Press 1982

[6] Bloch I. and Maître, H. Fuzzy Mathematical Morphologies: a comparative study, Télécom Paris 94D001

[7] Burillo, P., Frago, N. , Fuentes-González, R. Generation of Fuzzy Mathematical Morphologies. Mathware & Soft Computing Vol. VIII, n 1 31-46 (2001).

[8[De Baets, B. Fuzzy morphology: a logical approach, in Uncertainty Analysis, in B. Ayyub and M. Gupta (Eds.),

[9] Burusco, A.; Fuentes-González, R. Concept lattices defined from implication operators. Fuzzy Sets and Systems 114 431-436 (2000).

[10] Umbreit S. Formale Begriffsanalyse mit unscharfen Begriffen. Dissertation, Martin-Luther-Universität Halle-Wittenberg 1995.

[11] Bouchon-Meunier, B. Aggregation and fusion of imperfect information. Heidelberg: Physica-Verlag 1998

[12] Fuentes-González, R., Burillo, P. Y Frago, N. Técnicas de Morfología Matemática en el tratamiento de Sistemas Relacionales Difusos. ESTYLF2002. León (2002)

[13] Kohout L., Bandler W. Use of fuzzy relations in knowledge representation, adquisition and processing. In Zadeh L. and Kacprrzyk (Ed) Wiley 1992

[14] Soille P., Morphological Image Analysis. Principles and Applications. Springer 1999

[15] Pawlak, Z. Rough Sets. Kluwer 1991