Mining medical data to obtain fuzzy predicates

Taymi Ceruto[1], Orenia Lapeira[1], Annika Tonch[2, 3], Claudia Plant[2, 3], Rafael Espin[4,] Alejandro Rosete[1]

[1] Instituto Superior Politécnico "José Antonio Echeverría" (CUJAE), Havana, Cuba
{tceruto, olapeira,rosete}@ceis.cujae.edu.cu
[2] Helmholtz Zentrum, München, German Research Center for Environmental Health, Scientific Computing Research Unit, Germany
[3]Technische Universität München, Department of Informatics, Germany
{claudia.plant,annika.tonch}@ helmholtz-muenchen.de
[4] Universidad de Occidente (UDO), Sinaloa, México
rafaelespin@yahoo.com

Abstract. The collection of methods known as 'data mining' offers methodological and technical solutions to deal with the analysis of medical data and the construction of models. Medical data have a special status based upon their applicability to all people; their urgency (including life-or death); and a moral obligation to be used for beneficial purposes. Due to this reality, this article addresses the special features of data mining with medical data. Specifically, we will apply a recent data mining algorithm called FuzzyPred. It performs an unsupervised learning process to obtain a set of fuzzy predicates in a normal form, specifically conjunctive (CNF) and disjunctive normal form (DNF). Experimental studies in known medical datasets shows some examples of knowledge that can be obtained by using this method. Several kind of knowledge that was obtained by FuzzyPred in these databases cannot be obtained by other popular data mining techniques.

Keywords: Knowledge Discovery, Fuzzy Predicates, Medical Data

1 Introduction

Human medical data are at once the most rewarding and difficult of all biological data to mine and analyze [1]. Most of the people have some of their medical information collected in electronic form or at least in hard copy. This data may be collected from interviews with the patient, laboratory data, and the physician's observations and interpretations. These subjects generate vast volumes of data that can help to do a diagnosis, prognosis, and treatment of the patient and for that reason cannot be ignored. Thus, there is a need to develop methods for efficient mining in databases.

Data mining can be seen as a process that uses (novel) methods and tools to analyze large amounts of data. It has been applied with success to different fields of human endeavor, including marketing, banking, customer relationship management,

engineering and various areas of science [2]. However, its application to the analysis of medical data has gained growing interest. This is particularly true in practical applications in clinical medicine which may benefit from specific data mining approaches that are able to perform predictive modeling, to exploit the knowledge available in the clinical domain and to explain proposed decisions once the models are used to support clinical decisions [3].

In [4, 5] was proposed a singular way of extracting interesting knowledge from databases, called FuzzyPred. This approach restricts the representation of knowledge to a predicate in normal form. We believe that this kind of knowledge representation may be considered as a generalization, e.g. a conditional rule $A \rightarrow B$ is equivalent to the predicate $\neg A \vee B$. Moreover FuzzyPred can generate some interesting patterns that are impossible to be obtained by using other methods, e.g. (B) or (not B and C) or (D).

FuzzyPred integrates fuzzy set concepts and metaheuristic algorithms to search for logic predicates in a given data set [4]. The learning process is not supervised. We aim at evaluating how this technique can be applied on medical data and how they differ in terms of capabilities of discovering another kind of knowledge. As a result, this paper focuses on demonstrating its applicability in some medical datasets.

The paper presents a data mining study of medical data and it is organized as follows. Section 2 is an overview of knowledge discovery process and the related approaches with FuzzyPred. Section 3 is dedicated to explain FuzzyPred. Section 4 gives a brief overview of the implementation of FuzzyPred. A detailed description of the medical data we have used, the setup of all experiments and the results can be found in Section 5. Conclusions and proposal of future work are given in Section 6.

2 Preliminaries

Knowledge discovery in databases (KDD) is a non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns from large collections of data [2]. This process consists of several distinct steps and Data mining (DM) is the core step, which results in the discovery of hidden but useful knowledge from massive databases. DM tasks can be classified to tasks of description and prediction. The aim of description tasks is to find human-interpretable patterns and associations. On the other hand, the prediction task involves finding possible future values and/or distributions of attributes. Although the goals of them may overlap, the main distinction is that prediction requires the data to include a classification variable [6].

Over the last few years, the term data mining has been increasingly used in the medical literature [1, 3]. It is important in medical data mining, as well as in other kinds of data mining, to follow an established procedure of knowledge discovery, from problem specification to application of the results. The important issues are the iterative and interactive aspects of the process.

We list here some of the most commonly used data mining methods [6, 7]:

- **Decision tree** is a knowledge representation structure consisting of nodes and branches organized in the form of a tree such that, every internal non-leaf node is labeled with values of the attributes. It can be used to classify an unknown class data instance. Most current data mining suites include variants of C4.5 and CART decision tree induction algorithms; for instance Weka, Orange, KNIME.
- **Rule induction** is the process of extracting useful 'if -then' rules from data based on statistical significance. The antecedent (IF) contains one or more conditions about value of predictor attributes whereas the consequent (THEN) contains a prediction about the value of a goal attribute. It may be constructed from induced decision trees (as in the C4.5) or can be derived directly (Apriori algorithms).
- **Clustering** attempts to look for groups (clusters) of data items that have a strong similarity to other objects in the same group, but are the most dissimilar to objects in other groups. Popular clustering techniques include k-means clustering and expectation maximization (EM) clustering.

As shown below, the most conventional data mining algorithms identify the relationships among transactions using specific knowledge representation model (rules, trees, clusters). For that reason, the choice of the knowledge extraction method influences considerably the possible ways of knowledge representation. A final user is concerned with understanding and comprehending the extracted knowledge and that is where the form of knowledge representation plays an important role (depending on their potentialities and limitations).

Specifically, in order to obtain predicates (statement that may be true or false depending on the values of its variables), two main approaches are relevant from the literature:

- **Inductive Logic Programming** (ILP) [8]: ILP induces hypotheses from observations (examples) and synthesize new knowledge from experience. It needs a set of observations (positive and negative examples), background knowledge and hypothesis language.
- **Genetic Programming (GP)** [9]: GP is a branch of genetic algorithms. It is an automated method for creating a working computer program from a high-level problem statement of a problem. The learning is supervised. It exclusively uses genetic algorithms.

Recently, the fuzzy set theory [10] has been used more and more frequently in intelligent systems because of its simplicity and similarity to human reasoning. Specifically fuzzy mining methods for extracting implicit generalized knowledge from transactions stored are evolving into an important research area. It integrates fuzzy-set concepts and generalized data mining technologies to achieve this purpose. The mined patterns are expressed in linguistic terms, which are more natural and understandable for human beings. Several fuzzy learning algorithms (AprioriTid, Fuzzy ID3, Fuzzy C-Mean) for inducing patterns from given sets of data have been designed and used to good effect with specific domains [11, 12, 13, 14].

In general, several models of knowledge are impossible to be obtained by the previous methods. For instance, in a fuzzy database with variables A, B, C, and D, the following knowledge models may not be obtained:

- (A and B) or (not B and C)
- (A and B and not D)
- (B) or (not B and C) or (D)

The reason behind this is that the models of knowledge representation in the previous methods are limited. Some of these predicates may be part of the antecedent of a rule. However, they alone are not obtained as knowledge, and its quality is never calculated. It is significant to note that predicates can represent useful and valuable knowledge that describe the data from experts in various problem domains [15, 16, 17, 18, 19].

In a Boolean algebra every function can be represented by its Conjunctive Normal Form (CNF) and Disjunctive Normal Form (DNF) described by the binary linguistic values of true (1) and false (0). CNF is a normalization of a logical formula which is a conjunction of disjunctive clauses and DNF is a normalization of a logical formula which is a disjunction of conjunctive clauses [20]. It can be defined by the three primary operators of AND, OR, and NOT without losing any information from the precise combined concept. This implies that the normal forms in classic logic can be seen as general models to represent logic predicates.

Since there is no syntactical difference between formulas in fuzzy logic and formulas in two-valued logic, we can easily see that formulas in fuzzy logic can also be expressed in conjunctive and disjunctive normal form. In this case, they are valid expressions that hold as a matter of degree in the interval of [0, 1] which are bound by fuzzy normal forms known as fuzzy disjunctive and conjunctive normal forms [21]. Hence, the aim of our proposal is to obtain fuzzy predicates in normal form with high truth values, in medical databases.

3 FuzzyPred

Typically, a data mining algorithm constitutes some combination of the following three components: model, evaluation criteria and search algorithm [12]. The next sections describe FuzzyPred following these three components.

3.1 Model representation

Data in relational databases are stored in tables, where each row is the description of an object and each column is one characteristic/attribute of the object. In this case, a fuzzy transaction can contain more than one item corresponding to different labels of the same attribute, because it is possible for a single value in the table to match more than one label to a certain degree [22].

The process of converting an input value to a fuzzy value is called "fuzzification" and it may be done by using many of the available membership functions. Triangles, trapezoidal or left and right shoulder are commonly used because they give good results and their computation is simple [23]. **Fig. 1** shows three examples of a membership functions for the concepts young, mature and old in the interval 0 to 70 years.

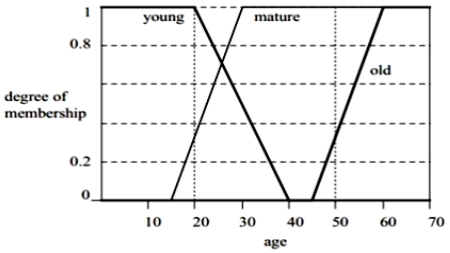

Fig. 1. Membership functions for the concepts young, mature and old [24]

The three functions in **Fig. 1** define the degree of membership of any given age in the sets of young, adult, and old ages. If a man is 20 years old, for example, his degree of membership in the set of young persons is 1.0, in the set of adults is 0.35, and in the set of old persons is 0.0. If another man is 50 years old, the degrees of membership are 0.0, 1.0, and 0.3 in the respective sets.

To compute this value we need to use the equation according to the type of membership function used. **Fig. 2** shows general equation for linear membership functions, defined by four points: a, b, c, d.

$$\mu(x)=\begin{cases} 0 & x \le a \\ \dfrac{x-a}{b-a} & a < x \le b \\ 1 & b < x \le c \\ \dfrac{d-x}{d-c} & c < x \le d \\ 0 & x > d \end{cases}$$

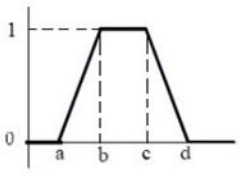

Fig. 2. General equation for linear membership functions

For example, the label young in **Fig. 1** represented by left-shoulder set has the equation 1. In this case the parameters are: a=b=0, c=20, d=40.

$$\text{young}(x) = \begin{cases} 1 & \text{if } 0 < age(x) \le 20 \\ \dfrac{40 - \text{age}(x)}{20} & \text{if } 20 < age(x) \le 40 \\ 0 & \text{if age}(x) > 40 \end{cases} \quad (1)$$

Predicates are commonly used to talk about the properties of objects, by defining the set of all objects that have some property in common [18]. In general, a predicate is a statement that may be true or false depending on the values of its variables. Nevertheless, in fuzzy logic, the strict true/false valuation of the predicate is replaced by a quantity interpreted as the degree of truth [25]. Fuzzy predicate may be a tree where each internal node may be an operator and each leaf is a fuzzy variable of the database. Besides, each linguistic variable can be associated with adverbs expressed in natural language called hedges. Hedges are terms that modify the shape of fuzzy sets. They have two main behaviors: reinforcement (such as "very") and weakening (such as "a little") [26].

In FuzzyPred each predicate is represented as a vector (SC, QC, NF) where the SC is a succession of clauses, the QC is the quantity of clauses and NF is the normal form. Each clause inside SC represents the attributes (fuzzy variable) and its values. We have used a positional encoding where the 'i' attribute is encoded in the 'i' gene used. When the integer value is '0', this attribute is not involved in the predicate, and when this part is different to '0' this attribute is part of the clause in the predicate. "1" indicates that the variable appear normal (x), "2" means that appears affected by the negation (1-x), and "3" indicates that the variable is associated to hedge "very", that implies that you need to square the value (x^2) when you will compute the fitness value. Figure 3 shows the scheme of a predicate for one example.

Finally, a predicate is coded in the following way:
Predicate = (SC, QC, NF)
SC= $C_1, C_{i,..} C_z$ where z = QC = quantity of clauses
C = $Var_1, Var_{,..} Var_y$ where y is the number of attributes in the dataset
QC= i where i>0
NF = {0 if CNF, 1 if DNF}

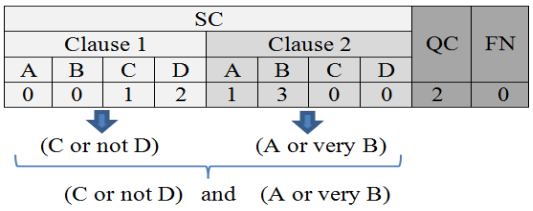

Fig. 3. Encoding of a predicate

3.2 Evaluation criteria

All techniques require a suitable measure to capture the correct model. In FuzzyPred there is only one measure to evaluate the quality of the fuzzy predicates: **F**uzzy **P**redicate **T**ruth **V**alue (FPTV) (see Equation 2-7 and Table 1) [15, 18, 27]. It depends on the number of clauses (z), variables (y) and records (x) of the data set. Table 1 gives a list of symbols used in this paper to define the formula.

Table 1. Symbols considered for the formula

Symbol	Definition
TV (var)	**T**ruth **V**alue of the **var**iable
TV (∧ clause)	**T**ruth **V**alue of the **Clause** in Conjunction
TV (∨ clause)	**T**ruth **V**alue of the **Clause** in Disjunction
TVC	**T**ruth **V**alue of the predicate in **C**NF for a single tuple
TVD	**T**ruth **V**alue of the predicate in **D**NF for a single tuple
FPTV	**F**uzzy **P**redicate **T**ruth **V**alue in all database

$$TV\ (var) = \begin{cases} var & if\ \ var = 1 \\ 1 - var & if\ \ var = 2 \\ var^2 & if\ \ var = 3 \end{cases} \quad (2)$$

$$TV\ (\wedge\ clause) = conj\ (TV(var)_1, .., TV(var)_Y) \quad (3)$$

$$TV\ (\vee\ clause) = disj\ (TV(var)_1, .., TV(var)_Y) \quad (4)$$

$$TVD\ = disj\ ((TV(\wedge\ clause)_1, .., (TV(\wedge\ clause)_Z) \quad (5)$$

$$TVC\ = conj\ ((TV(\vee\ clause)_1, .., (TV(\vee\ clause)_Z) \quad (6)$$

$$FPTV = \begin{cases} \forall_X TVC & if\ predicate\ is\ in\ CNF \\ \forall_X TVD & if\ predicate\ is\ in\ DNF \end{cases} \quad (7)$$

An example of how the predicate (Fig 3) can be evaluated in a small fuzzy database can be observed in Table 2. The FPTV is computed by using fuzzy logic operators. It is noteworthy that fuzzy logic does not give a unique definition of the classic operations as union or intersection. Different operators can be used (e.g. Min-Max [10], Compensatory [27-28]). In this case we use a compensatory fuzzy operator [27]: geometric mean to do a conjunction: $(x_1 * x_2 * ... * x^n)^{1/n}$ and its dual to do a disjunction: $1- ((1- x_1) (1- x_2)\ldots (1-x_n))^{1/n}$. In these operators, the associativity is excluded because it is incompatible with other desirable properties (idempotent, sensibility).

Table 2. Evaluation of the predicate step by step

Fuzzy dataset				TVvar (Equation 2)				TV (∨ clause) (Equation 3)		TVC (Eq 5)	FPTV
A	B	C	D	C	¬D	A	B^2	C∨ ¬D	A∨B^2	(C∨ ¬D) ∧(A ∨ B^2)	
0	0.4	0	0.2	0	0.8	0	0.64	0.55	0.4	0.47	
0.9	0.6	0	0.8	0	0.2	0.9	0.36	0.10	0.74	0.28	
0.8	0.4	0	0.6	0	0.4	0.8	0.64	0.22	0.73	0.4	0.5
0.5	1	1	0	1	1	0.5	1	1	1	1	
0.4	0.2	0.6	1	0.6	0	0.4	0.04	0.36	0.24	0.29	
1	0.6	0.2	0	0.2	1	1	0.36	1	1	1	

In Table 2, the first column is the original fuzzy data set. The second one represents the TV of all attributes involved in the predicate according to the operator or hedge applied (equation 2). For example in the case of ¬D the value is 1-D. Then it is necessary to compute the truth value of each clause in disjunction (equation 3). After, we calculate the TVC of complete predicate in each record (equation5). The last step consists of applying the universal quantifier in all records (conjunction of the values obtained in the previous column).

The value of FPTV is expressed by a real number in the interval [0, 1]. For that reason it may be interpreted like a fuzzy value, where '1'means that the statement is completely true, and '0' means that the statement is completely false, while values less than '1' but greater than '0 'represent that the statements are "partly true", to a given, quantifiable extent.

3.3 Search algorithm

In many cases the DM problem has been reduced to purely an optimization task: find the patterns that optimize the evaluation criteria. Metaheuristics represent a class of techniques to solve, approximately, hard combinatorial optimization problems. Some examples of metaheuristics are Hill Climbing (HC) and Genetic Algorithm (GA) [29-30]. Many successful applications have been reported for all of them. According to the "No Free Lunch" [31] it is impossible to say which is the best metaheuristic. It depends on the encoding, the objective function as well as the operators.

The global process in FuzzyPred tries to get predicates with high FPTV. The algorithm tries to maximize it as it is shown next:

```
BEGIN
   Predicate Set = Ø
   Initialize parameters
   IS = Generate random initial solutions
   Predicate Set = Predicate Set + IS
   REPEAT
   Pc = Generate new solution according to the metaheu-
   ristic selected
     If Pc is accepted
        IS = Pc
        Predicate Set=Predicate Set + Pc
   While stop condition is not verified
   Return Predicate Set
 END
```

The final result of the process is the concatenation of the predicates obtained by running the algorithm several times. Besides, FuzzyPred has included a phase of post-processing in order to improve the readability of the results.

Post-processing makes also possible to visualize and to store the extracted patterns. A standard data mining language or other standardization efforts will facilitate the systematic development of datamining solutions, to improve interoperability among multiple data mining systems and functions, and to use of data mining systems in industry and society [7].

Recent efforts in this direction include Predictive Model Markup Language (PMML) created by Data Mining Group [31]. PMML is an XML-based language that enables the definition and sharing of predictive models between applications. It is the de facto standard to represent predictive models. FuzzyPred exports the set of obtained predicates by using PMML.

FuzzyPred is a new way of obtaining knowledge that uses a different model and therefore it was necessary to adapt the original RuleSetModel (the nearest model) defined in PMML in order to create a new model called FuzzyPredicateModel. The labels "Header" and "DataDictionary" are maintained. In addition, FuzzyPredicateModel includes two fundamental labels: "MiningSchema" and "PredicateSet".

The original contributions of FuzzyPred are:

- The learning process is not supervised.
- The structure of the knowledge is not totally restricted, but it focuses only on fuzzy predicates.
- It represents a more flexible structure to allow each variable to take more than one value, and to facilitate the extraction of more general knowledge.
- Fuzzy logic contributes to the interpretability of the extracted predicates due to the use of a knowledge representation nearest to the expert.
- It is possible to use different fuzzy operators to calculate the truth value of the predicate (although compensatory is privileged because it has demonstrated to be highly efficient in the context of decision making).
- There is more than one search method (metaheuristics) available.

4 Implementation of FuzzyPred

Commercial data mining software is sometimes prohibitively expensive and the alternate open source data mining softwares are gaining popularity in both academia and in industrial applications. The Konstanz Information Miner (KNIME) [33] is a modular environment which enables easy visual assembly and interactive execution of a data pipeline. It is designed as a teaching, research and collaboration platform, which enables easy integration of new algorithms, data manipulation or visualization methods as new modules or nodes.

For that reason FuzzyPred was implemented in Java as a plugging in KNIME. Its user-friendly graphical workbench allows assembly of nodes for the entire analysis process. A flow usually starts with a node that reads in data from some data source.

Imported data is stored in an internal table-based format consisting of columns with a certain (extendable) data type (integer, string, etc.) and an arbitrary number of rows conforming to the column specifications. These data tables are sent along the connections to other nodes that modify, transform, model, or visualize the data.

Modifications can include handling of missing values, filtering of column or rows, oversampling, partitioning of the table into training and test data and many other operators. The node for transforming data (including the definition of membership functions) used Xfuzzy 3.0 [34]. Xfuzzy has been entirely programmed in Java and it is composed of several tools that cover the different stages of the fuzzy design. Specifically, we used Xfedit because the graphic interface of this tool allows the user to create and to publish the membership functions for each attribute using linguistic hedges as well as new fuzzy operators defined freely by the user.

The node for running the metaheuristics algorithms use an open source library called BICIAM [35]. It is a software tool for the resolution of combinatorial optimization problems by using generic algorithmic skeletons implemented in Java. It employs a unified model of metaheuristics algorithms, which allow us to define the problem only one time and execute the available algorithms many times. The node for visualizing the predicates obtained is supported in the tool SpaceTree [36].

The advantages are that each node stores its results permanently and thus workflow execution can easily be stopped at any node and resumed later on. Intermediate results can be inspected at any time and new nodes can be inserted and may use already created data without preceding nodes having to be re-executed. The data tables are stored together with the workflow structure and the nodes' settings.

5 Experiments

In this section we show the application of FuzzyPred to the analysis of public medical data, which comes from UC Irvine Machine Learning Repository (http://archive.ics.uci.edu/ml/). In particular we applied FuzzyPred to mine the datasets described in Table 3. These databases were selected taking into account their diversity, in terms of: pathology, number of attributes, types of attributes, total of tuples. In the third column of the table, (R / I / N) means (Real / Integer / Nominal).

In this study we aim at showing how the method could be suitably used to extract meaningful patterns that characterize the databases, highlighting interesting frequent associations. Membership functions for fuzzy sets can be defined in any number of ways [22]. The shape of the membership function used defines the fuzzy set and so the decision on which type to use is dependent on the purpose. Its choice is the subjective aspect of fuzzy logic, it allows the desired values to be interpreted appropriately [23].

Table 3 Datasets considered for the experimental study

Databases	Description	Attributes (R / I / N)	Records
BreastCancer Wisconsin (BC)	It contains cases from a study that was conducted at the University of Wisconsin Hospitals, Madison, about patients who had undergone surgery for breast cancer. The task is to determine if the detected tumor is benign or malignant.	(0 / 9 / 0)	699
Dermatology (D)	The differential diagnosis of erythemato-squamous diseases is a real problem in dermatology. Patients were evaluated clinically and histopathologically with 34 features.	(0 / 33 / 0)	366
Postoperative (P)	The goal of this database is to determine which patients in a postoperative recovery area should be sent to another area: I - Intensive Care Unit, S - go home and A- general hospital floor. Because hypothermia is a significant concern after surgery, the attributes correspond roughly to body temperature measurements.	(0 / 0 / 8)	90
Heart (HT)	This dataset is a heart disease database. The task is to detect the absence or presence of heart disease.	(1 / 12 / 0)	270
Mammographic (M)	The data was collected at the Institute of Radiology of the University Erlangen-Nuremberg between 2003 and 2006. This data set can be used to predict the severity (benign or malignant) of a mammographic mass lesion.	(0 / 5 / 0)	961

To develop demonstrative experiments, we extracted randomly three attributes from each database (in order to facilitate the interpretation), but you can also use more without limitation. The corresponding fuzzy sets related to the linguistic labels for each variable are specified through the corresponding membership functions. They were defined using mainly a partition with trapezoidal membership functions defined by a lower limit **a**, an upper limit **d**, a lower support limit **b**, and an upper support limit **c**, where $a < b < c < d$. The linguistic label used for each attribute to create the mining view was also taken by random choice (using the negation operator and hedges we can obtain the others in some way). In Table 4 the columns represent: D- database, LL- linguistic labels used, and a-d parameters for fuzzification.

Table 4. Definition of linguistic labels

D	LL	a	b	c	d
BreastCancer Wisconsin (BC)	Clump.Little	1.0	1.0	4.6	6.4
	CellShape.High	4.6	6.4	10.0	10.0
	Mitoses.Little	1.0	1.0	4.6	6.4
Dermatology (D)	Erythema.Little	0.0	0.0	1.2	1.8
	Eosinophils.High	1.2	1.79	3.0	3.0
	Scaling.High	1.2	1.79	3.0	3.0
Postoperative (P)	IntTemp.High	1 if x=high 0 if x=low or mid			
	SurfTemp.Mid	1 if x=mid 0 if x=low or high			
	OxySat.Excellent	1 if x=excellent 0 if x=good			
Heart (HT)	Age.Young	29	29	38	45
	ExerciseInduced.Few	0.0	0.0	0.2	0.4
	MaxHeartRate.High	71	150	202	202
Mammographic (M)	Age.Young	18	18	35	50
	Density.High	2.2	2.8	4.0	4.0
	Severity.Benign	1 if x=0 0 if x=1 (malignant)			

The following values have been considered in each experiment:

- Metaheuristics used for mining fuzzy predicates: HC, GA
- Genetic parameters: 20 individuals, 0.9 as crossover probability, 0.5 as mutation probability, single point crossover, uniform mutation.
- 30 repetitions were executed, each one with a maximum number of 500 iterations.
- Geometric Mean and its dual [26] were used to evaluate the predicates.

The algorithm returns several solutions in each run. Therefore, we show in Table 5 some representative solutions for each problem. The first column in Table 5 (Fuzzy Predicated Identifier, FPId) corresponds to an identifier associated with a predicate. The first part of the FPId identifies the corresponding database, e.g. D_2 is a predicate obtained from the database Dermatology (D). The second column is the predicate using the linguistic labels defined previously in Table 4. Finally, it appears the computation of FPTV.

Table 5. Examples of interesting fuzzy predicates

FPId	Predicate	FPTV
BC_1	CellShape.High or not Mitoses.Little	0,99
BC_2	Clump.Little or Mitoses.Little	0,95
BC_3	Clump.Little or CellShape.High or (not Mitoses.Little)	0,93
D_1	Erythema.Little or not Eosinophils.High or Scaling.High	1
D_2	Erythema.Little or (not Eosinophils.High) or (not Scaling.High)	1
D_3	not Erythema.Little or (not Eosinophils.High) or (not Scaling.High)	1
P_1	(not IntTemp.High) or (not SurfTemp.Mid) or	1

	(OxySat.Excellent)	
P_2	(very IntTemp.High) or (SurfTemp.Mid) or (OxySat.Excellent)	1
P_3	(not IntTemp.High) or (SurfTemp.Mid) or (not OxySat.Excellent)	0,87
HT_1	(not ExerciseInduced.Few) or (MaxHeartRate.High)	1
HT_2	(Age.Young) or (not MaxHeartRate.High)	0,98
HT_3	(Age.Young) or (not ExerciseInduced.Few) or not MaxHeartRate.High	0,97
M_1	(Severity.Mild) and (Density.High) and (not Density.High or not Severity.Mild)	1
M_2	(Density.High or not Severity.Mild)	1
M_3	(not Age.Young or not Density.High or Severity.Mild)	0,87

According to the results shown in Table 5 we can state the following conclusions about each database taking two predicates as examples:

- In the database Breast Cancer Wisconsin the Clump is Little or Mitoses is Little (BC_1). On the other hand the Mitoses is not Little or CellShape is High (BC_2).
- In the database Dermatology we can affirm with 100% of security that the Erythema is Little or Eosinophils is not High or Scaling is High (D_1).
- In the database Postoperative the Oxygen Saturation of patients is not Excellent or surface temperature in C is Mid (>= 36.5 and <= 35) or internal temperature in C is not High (< 37).
- In the database Heart the people are young or the maximum heart rate achieved is not High (HT_2)
- In the database Mammographic the Density is High or Severity is Malignant (M_2).

The objective of this experiment was to show the type of knowledge that can be obtained. From the obtained results we can observe that FuzzyPred generates fuzzy models with a good quality measure (maximum FPTV in some cases). All this fuzzy predicates lets us represent knowledge about patterns of interest in an explanatory and understandable form which can be used by the experts in each domain.

6 Conclusion

Fuzzy Mining is a useful technique to find patterns in data in the presence of imprecision, either because data are fuzzy in nature or because we must improve their semantics. It can be applied to create knowledge in rich medical environment. In this paper, we obtain good patterns through fuzzy predicates which represent dependence between items in the databases. The experimental results over five datasets highlighted the main potentials of the FuzzyPred, such as the opportunity to detect interesting relationships that could be implicitly hidden in the data.

Although the method worked well in these experiments, it is just a beginning. There is still much work to be done in this field. We will extend our experiments to other test data (more attributes and records) to extend the claims made in this paper. Besides, we are going to consider the possibility to automate the transformation of the fuzzy predicates in normal form to the way that the user desires (rules, groups) for better interpretation. It allows us to do some comparison with competing methods. Additionally other measures will be considered to evaluate the quality of results.

7 Acknowledgment

The authors would like to thank four anonymous reviewers for the helpful comments and suggestions.

8 References

1. Cios, K. J., William Moore, G.: Uniqueness of medical data mining. Artificial intelligence in medicine, vol. 26(1), pp. 1-24, (2002).
2. Fayyad U., Piatetsky-Shapiro G., Smyth P.: The KDD Process for Extracting Useful Knowledge, Communications of the ACM, vol. 39, pp. 27-34, (1996).
3. Bellazzi, R., Diomidous, M., Sarkar, I., Takabayashi, K., Ziegler, A., & McCray, A. Data analysis and data mining: current issues in biomedical informatics. Methods of information in medicine, vol. *50*(6), pp. 536, (2011).
4. Ceruto T., Lapeira O., Rosete A., Espin, R: Discovery of fuzzy predicates in database, Atlantis Press, Advances in Intelligent Systems Research, vol. 51. pp 45-54, ISSN 1951-6851, (2013).
5. Ceruto, T., Rosete, A., Espin R., Knowledge Discovery by Fuzzy Predicates. In Soft Computing for Business Intelligence, vol. 537, pp. 187-196, ISSN 1860-949X, Springer Berlin Heidelberg (2014).
6. Han J., Kamber M.: Data Mining: Concepts and Techniques (2nd edition), The Morgan Kaufmann Series in Data Management Systems, ISBN: 978-1-55860-901-3, pp. 1-14. (2006).
7. Berry M., Linoff M., Gordon S.: Data Mining Techniques, John Wiley & Sons, ISBN: 0-47L-47b4-3, pp. 11-40, (2004).
8. Muggleton S., DeRaedt L.: Inductive Logic Programming: Theory and methods. The Journal of Logic Programming, vol. 19-20, pp. 629-679, (1994).
9. Goldberg D., Koza J.: Genetic Programming Theory and Practice V, Springer Science+Business Media, ISBN-13: 978-0-387-76307-1, pp. 1-13, (2008).
10. Zadeh L.: Fuzzy Sets, Information Control, Vol. 8, pp-338-353, (1965).
11. Hong T., Lee Y.: An Overview of Mining Fuzzy Association Rules. Fuzzy Sets and Their Extensions: Representation, Aggregation and Models, Springer Berlin / Heidelberg, pp. 397-410, (2008).
12. Delgado M., Manín N., Martín-Bautista M.: Mining Fuzzy Association Rules: An Overview, Soft Computing for Information Processing and Analysis, Springer Berlin / Heidelberg. vol. 164, pp. 351-373, (2005).
13. Apolloni B., Zamponi G., Zanaboni, A.M. Learning fuzzy decision trees, Neural Networks vol 11, pp. 885–895, (1998).

14. Setnes M. and Kaymak U. Extended fuzzy c-means with volume prototypes and cluster merging, in Proc. EUFIT'98, Aachen, Germany, pp. 1360–1364, (1998).
15. Meschino, G., Espin R., Ballarin, V. A framework for tissue discrimination in Magnetic Resonance brain images based on predicates analysis and Compensatory Fuzzy Logic. IC-MED 2, No. X (1): 1-16, (2008).
16. Vanti A., Andrade, R. Administración Lógica: Un estudio de caso en una empresa de Comercio Exterior, Revista Base (Administração e Contabilidade) da UNISINOS, Vol 2(2): p. 69-77, (2005).
17. Delgado T., Delgado M. Evaluación del Índice de Alistamiento de IDEs en Iberoamérica y el Caribe a partir de un modelo de Lógica Difusa-Compensatoria, in Infraestructuras de datos espaciales en Iberoamérica y el Caribe, Casa editorial IDICT, pp 41-58, ISBN - 959-234-062-5, (2007).
18. Espín R., Fernandez E., Mazcorro G., Lecich M. A fuzzy approach to cooperative n-person games. European Journal of Operational Research, vol 176(3): p. 1735-1751, (2007).
19. Massone, H., et al., Evaluación de la peligrosidad de contaminación del agua subterránea mediante lógica difusa. Revista Argentina de Ingeniería (RADI), vol. 2(2), (2013).
20. Daňková M. Representation of logic formulas by normal forms, Kybernetika, Vol. 38, No. 6, pp 717-728, 2002.
21. Perfilieva, I. Normal forms for fuzzy logic functions in Multiple-Valued Logic, Proceedings (IEEE) of 33rd International Symposium , 2003.
22. Galindo J., Urrutia A., Piattini, M.: Fuzzy Databases: Modeling, Design and Implementation. Idea Group Publishing, ISBN 1-59140-325-1, pp. 341, (2006).
23. Mitsuishi T., Endou N. , Shidama Y. :The concept of fuzzy set and membership function and basic properties of fuzzy set operation. Journal of Formalized Mathematics, vol. 9, no. 2, pp. 315-356, ISSN 1426–2630, (2000).
24. Rojas, R.: Fuzzy Logic in Book Neutral Networks: A Systematic Introduction. ISBN 978-3-642-61068-4, pp. 502, Springer, (1996).
25. Cunningham, D.: A logical introduction to proof. New York: Springer. p. 29. ISBN 9781461436317, (2012).
26. Bouchon-Meunier B., Yao J.: Linguistic modifiers and imprecise categories. International Journal of Intelligent Systems Vol. 7, No. 1, pp. 25-36, (1992)
27. Espin R., Fernandez E., Mazcorro G. & et al. Compensatory Logic: A fuzzy normative model for decision making, Investigación Operacional, Vol. 27, No. 2, pp. 178-193, (2006),
28. Mizumoto M.: Pictorial Representactions of fuzzy conectives, Part II:cases of Compensatory operators and Self-dual operators, Fuzzy Sets and Systems, Vol. 32, pp. 45-79, (1989).
29. Talbi, E.: Metaheuristics: From Design to Implementation, Ed. John Wiley & Sons, ISBN 978-0-470-27858-1, pp. 18-29, (2009).
30. Blum C., Roli A.: Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison, *ACM Computing Surveys*, vol. 35(3), pp. 268–308, (2003).
31. Wolpert D., Macready W.: No Free Lunch Theorems for Optimization", IEEE Transactions on Evolutionary Computation, vol. 1, pp .67-82, (1997).
32. Data Mining Group, Welcome to DMG, Available: www.dmg.org, 4/6/2013
33. Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., & Wiswedel, B. KNIME: The Konstanz information miner Springer Berlin Heidelberg, pp. 319-326, (2008).
34. Xfuzzy Home Page, Fuzzy logic design tools, Available: www.imse-cnm.csic.es/Xfuzzy/
35. Fajardo J., Suarez A.: Algoritmo Multigenerador de Soluciones para la competencia y colaboración de generadores metaheurísticos, Revista Internacional de Investigación de Operaciones (RIIO), Colombia, ISSN 2145 - 9517, vol. 1 (0), pp-57-62, (2010).
36. SpaceTree, SpaceTree, Available: www.cs.umd.edu/hcil/spacetree/, 12/7/2013.