Knowledge discovery by Compensatory Fuzzy Logic predicates using a metaheuristic approach

Marlies Martínez Alonso[1], Rafael Alejandro Espin Andrade[1]

[1]"José Antonio Echeverría" Higher Technical Institute, Havana, Cuba.
marlies.martinez@gmail.com, espin@ind.cujae.edu.cu

Abstract

Compensatory Fuzzy Logic (CFL) is a logical system, which enables an optimal way for the modeling of knowledge. Its axiomatic character enables the work of natural language translation of logic, so it is used in knowledge discovery and decision-making. In this work we propose a general and flexible approach for knowledge discovery which allows obtaining different knowledge structure using a metaheuristic approach. The proposed method was tested by experimental analysis from a data set, using a tool developed in visual Prolog. The experimental results show some advantages of the proposed approach for representing patterns and trends from data.

Keywords: Compensatory Fuzzy Logic, Knowledge Discovery, Metaheuristic algorithms.

1. Introduction

The Knowledge Discovery in Databases (KDD) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in the data [1].

Data mining (DM) is the essential phase of KDD and it uses a number of techniques most frequently classified in two categories: Predictive (Direct) or Descriptive (Indirect) [2].

Currently there is a tendency to combine various techniques with the firm objective of resolving problems in which classic techniques did not quite well suited. Systems have followed this trend have increased their interest in the areas of Soft Computing (SC) and Computational Intelligence (CI).

Fuzzy Logic (FL) is a multivalent system (MS) that provides effective means for capturing the approximate and inexact nature of the real world by using its known interpretability from natural language. It has demonstrated a big potential to elaborate linguistic models, which make it very useful for solving real problems, by providing the proper schemes to improve communication with decision makers and experts [3].

MS has multiple advantages when modeling ambiguous or vague knowledge. However, these advantages still show certain difficulties when modeling knowledge in a natural way. For that reason, is frequent the use of isolated operators in the applications, or even the free combination between them along with an extra-logical resource called defuzzification [4].

Compensatory Fuzzy Logic (CFL) is an approach that proposes a new axiomatic theory for dealing with combinations of compensatory operators that enable simultaneous modeling of deductive and decision-making processes. CFL is distinguished because of its quality to generalize all the formulas of Bivalent Logic (BL). The CFL waives all compliance with the associative classical properties of conjunction and disjunction to achieve a closer representation of knowledge to human thought. Its ability to formalize reasoning, makes it possible to use it in situations requiring multi-criteria evaluations, also simultaneously taking into account predicates, which may even be contradictory [5][6].

In this paper we describe a method of knowledge discovery from obtaining CFL predicates. Many of the methods that currently exist for knowledge discovery focus on obtaining a given knowledge structure, (e.g. decision trees, rules, clusters) using a specific algorithm. For this reason, one of the motivations of this research is to propose a general and flexible method, which is not limited in regards to the structure of knowledge to be obtained, or the algorithm to be used. To achieve this flexibility, a declarative approach is intended to be used. It consists on separating and distinguishing the mechanisms to express the requirements of the mechanism used to satisfy them [7].

This paper begins by giving some basic insights of the CFL. Then it describes some of the basic notions of the proposed approach. Finally, experimental results show certain advantages of the proposed approach as a knowledge discovery tool.

2. Basic notions of Compensatory Fuzzy Logic

The main inconvenience of multivalent systems on the scope of knowledge modeling is that they do not achieve a generalization of BL formulas. This lack of flexibility sometimes causes a bad behavior in some interpretations of the logical variables. Another issue becomes evident when modeling decision-making problems. In these kinds of cases the deficiency is given by the associative character of the operators of conjunction and disjunction,

and the lack of sensibility to changes in the truth value of basic predicates when the truth value of compound predicates are calculated [5].

Fuzzy predicate logic has the same formulas as classical predicate logic (they are built from predicates of arbitrary arity using object variables, connectives &, →, truth constant 0 and quantifiers ∀ , ∃). A fuzzy predicate *f* with domain *X* is a function with images in [0, 1] where the image *f* (*x*) for each *x=a* belonging to *X* is called truth value of the predicate *f* for *x=a*. When you attribute directly the truth value of *f* for *x=a*, or using a membership function the predicate is called simple or basic. When a predicate *g* is a composition of basic predicates using the operators conjunction, disjunction, negation and implication, the predicate *g* is called a compound predicate.

The association is a main characteristic mostly of the operators used for aggregation. This characteristic is not good for data mining, because it equalizes the hierarchies of objectives and preferences. A very simple example which depicts this deficiency is:

We have a problem in which we want to create a model for selecting the outstanding student in a class. The fuzzy predicates are:

- *p= "good grades"*
- *q= "good attendance"*
- *r= "good class participation"*

It is known that the degree of "outstanding" of a student is given by the "good behavior" variable, which depends on good attendance and class participation ($p \wedge q$). The participation in school activities (*r*) counts as well. Therefore, the model to determine the outstanding student is represented as:

$$((p \wedge q) \wedge r)$$

However, the following expressions under the associative properties represent the same, and calculating the truth value of the compound predicate will give the same result for all of them.

$$((q \wedge r) \wedge p)$$
$$((p \wedge r) \wedge q)$$
$$((q \wedge r \wedge p)$$

The sensitivity is the ability to react to changes on the values of the predicates. This reaction generates different assessments and behaviors that affect the veracity of knowledge.

Finally, the compensation is the capacity to allow basic predicates to be compensated each other when the true values of the compound predicate are being calculated. Classical approaches of the Decision Theory include models such as additives, which accept compensation without limits. On the other hand multiplicative functional models and descriptive models accept a partial compensation, and are more favorable to decision making reasoning [5].

In this sense, CFL incorporates compensation axioms that make it a sensible system. That ensures that any variation in the truth values of basic predicates would modify the truth value of the compound predicate. This quality makes this logical approach behave more desirably to model the knowledge in a way closer to human reasoning, especially in decision-making processes.

Furthermore, multivalent systems generalize one or other BL properties, but none achieve generalizing completely. The CFL distinguishes itself from other multivalent logic systems, precisely because of its quality to generalize all formulas of BL properties with experimental analysis from a data set [6][8]. One of the most usual Compensatory Logics is the Geometric Mean Based Compensatory Logic (GMBCL).

2.1 GMBCL operators

In this section, GMBCL operators are shown, along with the way they are computed. In the following equations the truth values of the variables are inside the interval [0,1], where 1 represents absolute truth, and 0 the total denial of the truth.

The conjunction operator ($\wedge$ (and)) is calculated by the geometric mean of the truth value that takes the predicate of the analyzed variable. See equation 1. In this case *c* is the operator representing conjunction and x_n is the truth value of the variable *n*.

$$c(x_1, x_2, x_3, \ldots, x_n) = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_n} \quad (1)$$

The disjunction ($\vee$ (or)) is represented by the complement of the geometric mean of denials of truth values of the variables. It is calculated according to the equation 2. In this case, *d* is the disjunction operator and x_n is the truth value of the variable *n*.

$$d(x_1, x_2, \ldots, x_n) = 1 - \sqrt[n]{(1-x_1)\cdot(1-x_2)\cdot\ldots\cdot(1-x_n)} \quad (2)$$

The negation (¬) is calculated using the complement of the value of the variable denied. Equation 3 shows how to calculate it. In this case, *N* represents the negation operator and x_n represents the value of the variable *n*.

$$N(x_n) = 1 - x_n \quad (3)$$

The implication ($\rightarrow$) can be defined in two ways, shown in equations. In these equations, *i* represents the implication operator, *x* and *y* represent any pair of variables, *d* is the disjunction operator, *c* is the conjunction operator and *n* is the negation operator mentioned above.

$$i_1(x,y) = d(n(x),y)) \quad (4)$$

$$i_2(x,y) = d(n(x),c(x,y)) \quad (5)$$

The equivalence operator, also known as double implication ($\leftrightarrow$), is calculated by the conjunction of the implication and it's reciprocal. For any of the two variables (*x* and *y*), the equivalence can be defined, as shown in equation 6. In this equation *c* and *i* are the conjunction and implication operators presented above.

$$e(x,y) = c(i(x,y),i(y,x)) \quad (6)$$

The universal and existential quantifiers are calculated according to equations 7 and 8 respectively [8][9]. For any fuzzy predicate *p* in the universe *U*, universal and existential propositions are defined as the following:

$$\underset{x\in U}{\forall} p(x) = \underset{x\in U}{\wedge} p(x) \quad (7)$$

$$\underset{x\in U}{\exists} p(x) = \underset{x\in U}{\vee} p(x) \quad (8)$$

3. Knowledge Discovery Method

The method proposed in this research paper, uses a declarative approach. The declarative approach is based on the use of optimization methods and general purpose searches such as: Genetic Algorithms, Evolutionary Algorithms, Simulated Annealing, the Tabu Search and the classical methods of Artificial Intelligence such as the Stochastic Hill Climbing (SHC) [7].

To discover knowledge by obtaining CFL predicates, it is necessary to find good predicates in the space of the possible predicates. A predicate is considered good if it has a high truth value in the set of examples. Therefore, the problem is oriented to the use of a metaheuristic approach to optimize a function in the space of the predicates [10]. The proposed model consists of three fundamental elements that will be explained below:

- Knowledge representation in predicate form.
- Evaluation of predicates using CFL operators.
- Metaheuristic approach to perform searches.

3.1 Knowledge representation in predicate form

The language used in the proposed method is based on the predicate's logic. A predicate is an expression language that can connect with other expressions to form a sentence. The predicates have truth values depending on the terms. Therefore, a predicate may be true for one set of terms, and not for the rest. In this research, a variant of the FL is employed, and for this reason the predicates may have truth values within the interval [0, 1].

To represent the predicates we used general trees [36]. The general trees are defined as non-empty finite set T, of elements called nodes, such as:

- T contains an element distinguished R, called the root of T.
- The remaining elements of T form an ordered collection of disjoint trees T1, T2, .., Tn.

As it can be seen, this definition is recursive. Each sub-tree is itself a tree structure that follows the previous one.

In the model of the proposed method, the terminal nodes of a tree are the variables related to the problem, and the internal nodes of the tree are the operators (negation, $\neg$), (conjunction, $\wedge$), (disjunction, $\vee$), (implication, $\rightarrow$), (double implication or equivalence ($\leftrightarrow$)) of the CFL.

3.2 Evaluation of predicates using CFL operators

The main objective of the proposed method is to obtain CFL predicates with high evaluation. To evaluate predicates we have taken into account the following characteristics:

1. The truth value that acquires the predicate in the dataset.

This characteristic is important because obtaining predicates with high truth value (close to 1), means that, the knowledge obtained has a high value of certainty. In the predicate logic, universal and existential quantifiers are frequently used. The universal quantifier determines whether a formula (predicate) is true for all values within the domain. The existential quantifier indicates of a formula is true for any values within the domain. In the proposal method we use the CFL universal quantifier (see equation 7) to calculate the truth value that acquires a predicate in the dataset.

2. Penalize predicates that have repeated variables (to avoid obvious predicates such as $p \vee p \rightarrow p$).

The second characteristic is important because it prevents obtaining evident predicates, which acquire high truth value, but at the same time provide little new knowledge. To achieve this goal we defined penalize with

cero evaluation (0) those predicates with repeated variables.

3. Guide the search to obtain specific knowledge structures if needed. (e.g. association rules, classification rules, clustering, supervised learning, etc.)

The third characteristic is used to obtain different knowledge structures. In this case, a better evaluation is given to predicates that have the desired knowledge structure.

4. Obtain predicates with the largest amount of variables involved in the examples.

The fourth characteristic is defined in order to obtain more correlations between the variables analyzed and therefore more knowledge. In this case, a better evaluation is given to predicates that have more variables.

5. Obtain small predicates.

The fifth characteristic is a natural interest from the point of view of knowledge engineering, because small predicates are easier to interpret than large and complex predicates.

Equation 9 shows the function that computed the evaluation of CFL predicates:

$$E = \frac{T \cdot DV}{SC \cdot (SC - SSC + 1)} \quad (9)$$

Where:

- Evaluation (E): evaluation of the predicate.
- Truth value (T): truth value of the predicate in the set of data examples.
- Number of different variables (DV): number of different variables present in the tree.
- Size with constant (SC): number of nodes (terminal and internal) having the tree counting constants.
- Size without constant (SSC): number of nodes (terminal and internal) that has the tree without counting the constants.

Equation 9 represents the objective function to be optimized to search CFL predicates. This function aims to find predicates with high truth values, small, and with the most variables involved in the examples. In this equation, the evaluation is directly proportional to the truth value and the number of variables to be used, and inversely proportional to the length.

3.3 Metaheuristic approach to perform searches.

The metaheuristic approach uses two basic mechanisms: the evaluation mechanism, and the search mechanism. These mechanisms are implemented both separately and independently. This feature facilitates the scalability and flexibility of the proposed method because it allows modifying the search algorithm without interfering with the requirements to be met by the predicates. The method proposed in this study uses three basic and independent components:

- Mutations (generate new solutions from others and enable to obtain different knowledge structures).
- Evaluation (mentioned in the previous section).
- Search algorithm.

In the calculation of neighborhood each predicate defines a possible way to change. For each mutation, a predicate generates a random number that decides which of the possible operators should be applied. Considered in the neighborhood of any given CFLk predicate all CFLh predicates that derives from it, in the application of the following operators:

Operators 1 (generals)

- A terminal node is replaced by another terminal node, chosen randomly.
- An internal node is replaced by another internal node, taken at random.
- A sub-tree is replaced by a terminal node, both selected randomly.
- A terminal node is replaced by a sub-tree, both picked at random.

General operators allow obtaining predicates with different knowledge structures.

To obtain specific knowledge structures we defined the following operators:

Operators 2:

Operators to obtain classification rules:

- Establish the implication operator (→) as the root node.
- Establish the variable representing the class as a consequent of the rule.
- Using the conjunction operator (∧) for relating the variables that are in the rule antecedent.
- Only mutate the antecedent of the rule using the Operators 1, with the exception of mutation number 2, because no change of logic connective always is the conjunction (∧).

Operators for supervised learning:

- Establish the double implication operator (↔) as the root node.
- Establish the variable representing the class in one of the ends of the equivalence.
- Establish the conjunction (∧) as the logical connective between variables.
- Mutate the extreme of equivalence in which the variable that represents the class does not appear. For that mutation we can use the Operators 1, with the exception of mutation number 2, because no change of logic connective always is in the conjunction (∧).

Operators to obtain clustering:

1. Establish the disjunction operator (∧) as the root node of the tree.
2. Establish clauses in conjunction as sub-trees.
3. Only mutate the clauses in conjunction using Operators 1, with the exception of mutation number 2, because no change of logic connective always is conjunction (∧).

In the search algorithm the initial solution is generated depending of the knowledge structure that we want to obtain. If we want to obtain randomly different knowledge structures (e.g. classification rules, clustering, etc.) we use the mutation Operators 1 and if we want direct the search to obtain more specific structures we use Operator 2. The search ends when we find a tree with desired truth value or when reaching the predefined number of iterations.

4. Experiments

In order to prove the proposed method, we implemented an experimental tool that will allow us to make experiments using real data. As mentioned above we used general trees to represent knowledge. These structures are completely recursive and for that reason we have used Visual Prolog to implement this tool. This is an object-oriented programming language, that like Prolog, it has a high ability to define concepts in terms of themselves [11].

In these experiments we defined use Stochastic Hill Climbing (SHC) as the metaheuristic algorithm. To obtain each one of the predicates, 800 iterations of SHC were employed. This metaheuristic algorithm is simple and useful. However, the emphasis on SHC is only the starting point for future research. The experiments were oriented to obtain association rules, general predicates, classification rules, and supervised learning. The primary metric used to assess the quality of the results is the evaluation that acquires the predicates from the data set.

4.1 Data Set

To perform the experiments we used two cases of study. The first one was performed using a database with 1728 records extracted from the UCI Machine Learning Repository; this database has been used in several studies to classify cars according to their characteristics. The second case of study was performed using a database with 1989 records of diabetes patients. This database was extracted from data files provided by the Jaruco Health Center, Mayabeque Province, Cuba.

4.2 Case Study: Car

The car database contains 7 variables describing characteristics associated with each one of the cars. These variables are: the buying price, price of maintenance, number of doors, number of people (capacity in terms of persons to carry), the size of luggage boot, estimated safety of the car and car acceptability. The proposed method works with a variant of Fuzzy Logic. Therefore, it was necessary define a degree of membership for each variable with respect to previously defined fuzzy sets.

4.2.1 Experimental Results

To obtain predicates and classification rules through supervised learning, we had to use 1200 data records to train and create models. Then we added 584 complementary records to test these models. By using this testing method it was possible to verify that the model is valid when predicting new data.

The following results show the obtained predicates with the best evaluation. The variables V, VE and VP represent the truth value, truth value in training, and truth value in the testing respectively.

General predicates:

((((BuyingPrice = low) and (Safety = high)) and (((NumPeople = high) and (NumDoors = high)) and ((MaintenancePrice = low) and (LuggageBootSize = big))))) → (Acceptability = good))

V=0.9356

((((LuggageBootSize = big) and (BuyingPrice = low)) and (Acceptability = good))) ((Safety = high) and ((NumPeople = high) or (MaintenancePrice = low))))

V=0.9238

Association Rules:

(((BuyingPrice = low) and ((Safety = high) and (Acceptability = good))) and ((NumPeople = high) and

(MaintenancePrice = low))))) → ((NumDoors = high) or (LuggageBootSize = big)))

V=0.9128

((((BuyingPrice = low) and (Acceptability = good)) and (Safety = high))) → ((LuggageBootSize = big) or ((NumPeople = high) or (MaintenancePrice = low))))

V= 0.9125

((Acceptability = good)) → (Safety = high))

V=0.8782

Classification Rules:

((((BuyingPrice = low) and (Safety = high)) and (((NumPeople = high) and (NumDoors = high)) and ((MaintenancePrice = low) and (LuggageBootSize = big))))) → *(*Acceptability = good))

VE=0.9356
VP=0.9041

(((LuggageBootSize = big) and (((NumPeople = high) and (MaintenancePrice = low)) and (((BuyingPrice = low) and (NumDoors = high)) and (Safety = high))))) → (Acceptability = good))

VE=0.9345
VP=0.8764

(((((NumPeople = high) and ((LuggageBootSize = big) and (MaintenancePrice = low))) and (Safety = high)) and (BuyingPrice = low))) → (Acceptability = good))

VE=0.9139
VP=0.8121

Supervised learning:

(((BuyingPrice = low) and (MaintenancePrice = low)) ↔ (Acceptability = good))

VE=0.9054
VP=0.7923

(((BuyingPrice = low) and (LuggageBootSize = big)) ↔ (Acceptability = good))

VE=0.8014
VP=0.7425

(((BuyingPrice = low) and (NumPeople = high)) ↔ (Acceptability = good))

VE=0.7656
VP=0.6452

((BuyingPrice = low) ↔ (Acceptability = good))

VE=0.7041
VP=0.6321
((BuyingPrice = low) ↔ (Acceptability = good))

VE=0.7041
VP=0.6321

4.2.2 Analysis of results

The majority of predicates obtained in these experiments have a truth value above 0.70. Knowledge obtained through this experiment expresses a set of relationships between the variables used, which are listed below:

- Relationship between low buying price, and low price of the maintenance with the car acceptability.
- Relationship between high number of people, and high number of doors with the car acceptability.
- Relationship between high security and car acceptability.
- The best predicates (higher truth value) are those involving the values: high number of people, more number of doors, and big size of luggage boot, low buying price and low price of the maintenance, high estimated safety and good acceptability.
- The predicates with low truth value were obtained using supervised learning. This can be due to the fact that equivalences are more demanding structures.

The car database is derived from a hierarchical model of decision developed by [12]. The model evaluates cars according to certain concepts. The general description of this model defines a car with high level of acceptance if it meets the following characteristics:

- The buying price and price of the maintenance are medium or low.
- The technical features (number of people, number of door, size of luggage boot and estimated safety of the car) have high values (e.g. acceptable, good or very good).

Many of the predicates obtained in these experiments express relationships between variables similar to this model.

In order to compare some results of the proposed method with another method that has a similar purpose, we have to compare some classification rules obtained in this experiment; along with other classification rules

obtained using the JRip classification algorithm of the WEKA tool.

WEKA (Waikato Environment for Knowledge Analysis) is a framework for experimental data analysis that allows applying, analyzing and evaluating relevant techniques of data analysis, mainly those coming from machine learning, over any user data set [13].

Making this comparison we found several similarities between the results obtained by the JRip algorithm and some of the classification rules obtained by the proposed method. Some of these similarities are:

Weka JRiP:
(Safety = high) and (LuggageBootSize = medium) and (NumPeople = 4) and (BuyingPrice = low) and (NumDoors = 4)) → (Acceptability=vgood)

CFL predicate:
((((BuyingPrice = low) and (Safety = high)) and (((NumPeople = high) and (NumDoors = high)) and ((MaintenancePrice = low) and (LuggageBootSize = big))))) → (Acceptability = good))

V= 0.9356

****

Weka JRiP:
(Safety = high) and (BuyingPrice = medium) and (LuggageBootSize = big) and (MaintenancePrice = low) and (NumPeople = 4)) → (Acceptability = vgood)

CFL predicate:
(((((NumPeople = high) and ((LuggageBootSize = Big) and (MaintenancePrice = low))) and (Safety = high)) and (BuyingPrice = low))) → (Acceptability = good))

V= 0.9139

****
Weka JRiP:
(Safety = high) and (LuggageBootSize = medium) and (NumPeople = 4) and (BuyingPrice = low) and (NumDoors = 4)) → (Acceptability = vgood)

CFL predicate:
(((NumDoors = high) and ((BuyingPrice = low) and ((Safety = high) and (NumPeople = high))))) → (Acceptability = good))
V= 0.8236

****
Weka JRiP:
(Safety = high) and (BuyingPrice = low) and (NumPeople = 4)) → (Acceptability = good)

CFL predicate:

(((NumPeople = high) and ((BuyingPrice = low) and (Safety = high)))) → (Acceptability = good))

V= 0.7526

4.3 Case Study: Diabetes

The diabetes database contains 7 variables describing characteristics associated with each patient. These variables are: Race, Hypertension, Body Mass Index (BMI), cardiovascular and/or cerebral vascular accident antecedents (both known for the expression: "Antecedents"), Sex, Classification of diabetes (Classification). As in the previous experiment, each variable is assigned a membership value with respect to a fuzzy set.

4.3.1 Experimental Results

To obtain predicates and classification rules through supervised learning, we used 1400 data records to train and create models. An additional 589 were employed to test these models.
The following results are the obtained predicates which had better evaluation. The variables V, VE and VP represent the truth value, truth value in training and truth value in the testing respectively.

General predicates:

(((Race = white) and ((Antecedents = true) and (Age = advanced)))) → ((Classification = diabetic) or (BMI = high)))

V= 0.9256

((Race = white) or ((((Antecedents = true) and (Age = advanced)) and (Hypertension = true))) → (((Classifications = diabetic) or (BMI = high))) → (Sex = male))))

V= 0.9205

((Antecedents = true)) → ((((Hypertension = true) and (BMI = high)) and (Age = advanced)) or (Race = white)))

V= 0.8603

((Antecedents = true)) → (Age = advanced))

V= 0.8200

Association Rules:

(((Classification = diabetic) and (Antecedents = true))) →(Sex =
male))
V= 0.9532

((((Race = white) and (Classification = diabetic)) and (Age = advanced))) → (BMI = high))
V= 0.9189

(((Antecedents = true) and (Age = advanced))) → (((Race =
white) and (Sex = male)) or ((BMI = high) or (Hypertension = true))))
V= 0.9166

((Antecedents = true)) → ((Race = white) or (Age = advanced)))
V= 0.9125

(Age = advanced))) → (Hypertension = true)
V= 0.9036

Classification Rules:

(((((Sex = male) and (Race = white)) and (Hypertension = true)) and (Age = advanced))) → (Classification = diabetic))
V= 0.9741
VP= 0.8874

(((Antecedents = true) and ((BMI = high) and (Hypertension = true)))) → (Classification = diabetic))
V= 0.9014
VP= 0.8452

((((Age = advanced) and (Antecedents = true)) and (((Hypertension = true) and (Sex = male)) and (Race = white)))) → (Classification = diabetic))
V= 0.8907
VP= 0.8147

((((BMI = high) and ((Antecedents = true) and (Hypertension = true))) and (Age = advanced))) → (Classification = diabetic))
V= 0.8713
VP= 0.7875

(((Antecedents = true) and (Age = advanced))) → (Classification = diabetic))
V= 0.8347
VP= 0.8041

Supervised learning:

((Age = advanced) → (Classification = diabetic))
V= 0.8941
VP= 0.8023

(Race = white) and (Age = advanced)) → (Classification = diabetic))
V= 0.8632
VP= 0.7698
(((Antecedents = true) and (Age = advanced)) → (Classification = diabetic))
V= 0.8574
VP= 0.7758

(((Hypertension = true) and (Age = advanced)) → (Classification = diabetic))
V= 0.7041
VP= 0.6321

(((Antecedents = true) and (Age = advanced)) → (Classification = diabetic))
V= 0.7014
VP= 0.6142

(((Sex = male) and (Age = advanced)) → (Classification = diabetic))
V= 0.6912
VP= 0.5423

4.3.2 Analysis of results

In this experiment the majority of the predicates have truth value above 0.80. Knowledge obtained through this experiment expresses a set of relationships between the variables used, which are the following:

- Relationship between advanced age, hypertension and obesity.
- Relationship between cardiovascular and/or cerebral vascular accident and elderly people.
- Relationship between cardiovascular and/or cerebral vascular accident, hypertension, obesity and have diabetes.
- Relationship between the male sex and the presence of diabetes.
- Relationship between white race and the presence of diabetes.

In order to evaluate the quality of knowledge obtained through this experiment we made an extensive research about the real characteristics of diabetes. According to investigations, obesity increases the risk of diabetes and the risk of developing hypertension. Diabetes and hypertension commonly coexist; the appearance of both is common in elderly people [14][15]. The predicates obtained in this experiment reflect the real characteristics of diabetes, or at least show many similarities to the characteristics of the illness. The risk of suffering from diabetes is not dependent on sex and race; virtually any person can have diabetes [16]. Therefore, the influence of sex and race in the diabetes data set can be considered a novel discovery.

In this experiment we also did comparisons between some predicates obtained by the proposed method, with other results obtained using a method of similar purpose. In this case, we compare some classification rules obtained through this experiment with other classification rules obtained using JRip and Weka PART classification algorithm of the WEKA tool.

In this comparison we also found several similarities between the results obtained by the WEKA algorithms

and some of the classification rules obtained by employing the proposed method. Some of these similarities are:

Weka JRiP:
(Antecedent = true) and (BMI = high) and (Hypertension = true) → Classification = diabetic

CFL predicate:
(((Antecedents = true) and ((BMI = high) and (Hypertension = true)))) → (Classification = diabetic))

V= 0.9014
****

Weka JRiP:
(Antecedent = true) and (BMI = high) and (Hypertension = true) and (Age >= 58) and (Age <= 78)) → Classification = diabetic

CFL predicate:
((((BMI = high) and ((Antecedents = true) and (Hypertension = true))) and (Age = advanced))) → (Classification = diabetic))

V= 0.8713
****

Weka PART:
Sex = male AND
Hypertension = true AND
Race = white AND
Age >65 AND
Age <= 71: Classification = diabetic

CFL predicate:
(((((Sex = male) and (Race = white)) and (Hypertension = true)) and (Age = advanced))) → (Classification = diabetic))

V= 0.9741
****
Weka PART:
Race = white AND
Sex = male AND
Age >63 AND
Age <= 66: Classification = diabetic

CFL predicate:
(((Race = white) and ((Sex = male) and (Age = advanced))))
→ (Classification = diabetic))

V= 0.7173
****
Weka PART:
Antecedents = true AND
Age >62 AND
Age <= 73: Classification = diabetic

CFL predicate:
(((Antecedents = true) and (Age = advanced))) → (Classification = diabetic))

V= 0.8045

5. Conclusions

The approach proposed in this research allows obtaining different forms of knowledge structure, and it is not limited to using a specific algorithm. This proposal does not replace existing methods for knowledge discovery, but it is general way of extracting useful knowledge from data. The experimental results allows us to suppose that this approach has similar results to the ones obtained the particular methods. Then, its generality is an opportunity for improving systemic analysis and compound Data Mining problems, without losing quality in the solution of the particular problems. In the future we intend to make new experiments and comparisons between each type of problem, with other algorithms of similar purpose. In addition we intend to improve the tool, to be able to obtain more structures of knowledge, and combine the use of different metaheuristic algorithms. We also intend to apply approximate reasoning with the direct use of Prolog.

References

[1] U. Fayyad, G. Shapiro, and P. Smyth. *Advances in Knowledge Discovery and Data Mining*. Pearson Education, 1996.

[2] M. Berry and L. Gordon. *Data Mining Techniques*, Second Edition. John Wiley and Sons, 2004.

[3] L. Hongxing and Y. C. Vincent. *Fuzzy Sets and Fuzzy Decision-Making*. N. W. Boca Raton: CRC Press, 1995.

[4] H. J. Zimmermann. *Fuzzy Set Theory and its applications*. Kluwer Academic Publishers, 1996.

[5] R. A. Espin and E. G. Fernandez. "Compensatory Fuzzy logic: A platform for reasoning and knowledge representation in a muticriterio decision". In *Multicriteria for Decision Making: Methods and Applications*, pages 338-349. Editorial Plaza and Valdes / Publishing, 2009.

[6] R. A. Espín, G. T. Mazcorro, and E. G. Fernandez. "Normative considerations of compensatory fuzzy logic". In *Evaluation and Enhancement of Spatial Data Infrastructure for Sustainable Development in Latin America and the Caribbean*, pages 28-40. IDICT Edition, 2007.

[7] T. Lin and P. Eades. *Integration of Declarative and Algorithmic Approaches for Layout Creation*, Proceedings of Graph Drawing94. Springer-Verlag, 1994.

[8] S. Kleene. *Introduction to Metamathematics*. Princeton, 1952.

[9] R. A. Espín, G. T. Mazcorro, and E.G. Fernandez. *Towards a semantic normative approach Decision Making: Theoretical and experimental nexus of Compensatory Fuzzy Logic, the theory of expected utility and prospect theory*. IDICT Edition, 2007.

[10] R. Korf and K. E. *"Search", Encyclopedia of Artificial Intelligence*. Wiley Interscience Publications, 1993.

[11] M. Bramer. *Logic Programming with Prolog*. British Library, 2005.

[12] M. Bohanec and V. Rajkovic. "Knowledge acquisition and explanation for multiattribute decision making". In *8th Intl Workshop on Expert Systems and their Applications*, 1988.

[13] I. H. Witten and E. Frank. *Data Mining Practical Machine Tools abd Techniques.* Morgan Kaufmann Publishers, 2005.

[14] F. Messerli, D. Bell, and G. Bakris. *Body Weight Changes with Beta-Blocker Use: Results from GEMINI*. American Journal of Medicine, 120(7):610-615 Jul, 2007

[15] F. Contreras, M. Rivera, J. Vásquez, C. J. Yánez, M. A. Parte, and M. Velasco. *Diabetes and hypertension clinical and therapeutic aspects*. Scielo, 19(1):3-62, 2000.

[16] T. Zegarra, G. Guillermo, C. Caceres, and M. Lenibet. *Sociodemographic and clinical characteristics of type 2 diabetic patients with community-acquired infections admitted Medicine services Cayetano Heredia National Hospital*. Scielo, 11(3):3-62, 2000.