

# Quality measures for fuzzy predicates in conjunctive and disjunctive normal forms

## Medidas de calidad para predicados difusos en forma normal conjuntiva y disyuntiva

T. Ceruto<sup>1</sup>, O. Lapeira<sup>2</sup> and A. Rosete<sup>3</sup>

### ABSTRACT

Association rule mining is a very popular data mining technique. Rules in this technique are often used to identify and represent dependencies between attributes in databases. Specifically, fuzzy association rules are rules that use the concepts of fuzzy sets and can be considered as a special case of fuzzy predicates. Many quality measures have been defined for fuzzy association rules, but all consider a specific structure: antecedent and consequence. In the case of fuzzy predicates in the normal form (i.e., conjunctive or disjunctive), it is necessary to define different quality measures that do not consider the structure as an antecedent or a consequence. The only available measure for this scenario is the fuzzy predicate truth value (FPTV), which has serious limitations. The evaluation of fuzzy predicates in the normal form through appropriate quality measures has not yet been clearly defined in the literature. Thus, we propose several quality measures specifically for fuzzy predicates in the conjunctive (CNF) and disjunctive (DNF) normal forms. Experimental studies illustrate the use of the proposed measures and allow some general conclusions about each measure.

**Keywords:** data mining, fuzzy predicate, quality measures, conjunctive and disjunctive normal forms.

### RESUMEN

La extracción de las reglas de asociación es una técnica de minería de datos muy popular, las cuales son utilizadas a menudo para identificar y representar dependencias entre atributos en bases de datos. Específicamente, las reglas de asociación difusas utilizan conceptos de conjuntos difusos y pueden ser vistas como un caso especial de predicados difusos. Muchas medidas de calidad han sido definidas para reglas de asociación difusa, pero todas consideran la estructura específica de reglas: antecedente y consecuente.

En el caso general de predicados difusos en forma normal (conjuntiva o disyuntiva), es necesario definir diferentes medidas de calidad que no estén en función de antecedente y consecuente, puesto que la única medida disponible para ello, es el valor de verdad para predicados difusos (FPTV) y tiene serias limitaciones. La evaluación de un predicado difuso en forma normal, a través de medidas adecuadas de calidad no ha sido todavía claramente definida por otros autores. Por esa razón, en este trabajo se proponen varias medidas de calidad para los predicados difusos, en formas normal conjuntiva o disyuntiva. Los experimentos demuestran el uso que se le puede dar a las métricas propuestas y permiten llegar a conclusiones generales de cada una de ellas.

**Palabras clave:** minería de datos, predicados difusos, medidas de calidad, forma normal conjuntiva y disyuntiva.

Received: January 20th 2014

Accepted: May 5th 2014

### Introduction

Knowledge discovery, whose objective is to obtain useful knowledge from data, is recognized as a basic necessity. The theory of fuzzy sets can certainly aid in the data-mining process to reach this goal. Fuzzy sets handle numerical values better than existing methods because fuzzy sets soften the effect of sharp bound-

aries (Zadeh, 1965; Fayyad, Piatetsky-Shapiro & Smyth, 1996; Duarte, 1999; Han & Kamber, 2006; Venugopal, Srinivasa & Patnaik, 2009).

Many techniques used in datasets have their corresponding "fuzzy version." For instance, fuzzy association rules described by the natural language are well suited for human comprehension and help to increase the flexibility for supporting users in making deci-

<sup>1</sup> Taymi Ceruto. Ingeniero Informático, Master en Ciencias, Instituto Superior Politécnico "José Antonio Echeverría", Cuba. Affiliation: Cujae, Cuba. E-mail: tceruto@ceis.cujae.edu.cu

<sup>2</sup> Orenia Lapeira. Ingeniero Informático, Instituto Superior Politécnico "José Antonio Echeverría", Cuba. Affiliation: Cujae, Cuba. E-mail: olapeira@ceis.cujae.edu.cu

<sup>3</sup> Alejandro Rosete Suárez. Ingeniero Informático, Doctor en Ciencias, Instituto Su-

perior Politécnico "José Antonio Echeverría", Cuba. Affiliation: Cujae, Cuba. E-mail: rosete@ceis.cujae.edu.cu

**How to cite:** Rodríguez-Gutiérrez, J. A., & Aristizabal-Ochoa, J. D. (2014). Biaxial bending of slender HSC columns and tubes filled with concrete under short- and long-term loads: II) Verification. *Ingeniería e Investigación*, 34(3), 63-69.

sions (Delgado, Marín, Sánchez & Vila, 2003). Fuzzy clustering generally provides a more suitable partition of a set of objects than does classical clustering (Han & Kamber, 2006).

The discovery of fuzzy predicates in conjunctive (CNF) and disjunctive normal form (DNF) provides a convenient and effective general way to identify and represent certain dependencies among items in fuzzy transactions (Ceruto, Lapeira, Rosete & Espin, 2013). We believe that fuzzy predicates in CNF and DNF can be generalized somewhat because they produce patterns that the classic methods cannot obtain; we can also use it to generate an equivalent pattern. For instance, in classic logic  $\neg A \vee B$  is equivalent to a conditional rule  $A \rightarrow B$  (Bruno, 1998; Trillas, 2009).

It is worth clarifying that some logical expressions that are equivalent in classical logic may have different truth values in fuzzy logic. In addition, the truth value of a formula depends on the type of fuzzy operator used. This implies that two formulas that are equivalent in classical logic are only approximately equivalent (i.e., only to a certain degree) in fuzzy logic. For example, it has been stated that the Axioms of Kleene Algebra are more true than false in Compensatory Fuzzy Logic (Espin, Fernandez, Mazcorro, Marx-Gómez & Lecich, 2006).

Fuzzy predicate mining is a task that can be faced as an optimization problem. You can combine fuzzy set concepts and higher-level procedures (i.e., metaheuristics) to find and generate automatically good fuzzy predicates. This learning process is not supervised.

All techniques require a suitable measure to evaluate the model correctly. When mining fuzzy predicates, only one quality measure is of value: the **Fuzzy Predicate Truth Value (FPTV)** (Ceruto, Lapeira, Rosete & Espin, 2013). Although the FPTV is not robust to outliers, it penalizes the presence of zeros strongly (i.e., veto criteria). This property introduces the necessary capacity of restriction in compensation when certain goals are not fully satisfied (Espin, Fernandez, Mazcorro, Marx-Gómez & Lecich, 2006). The use of a universal quantifier also restricts the final output, which will be determined by the type of fuzzy logic operator selected. As a result, we can conclude that other formulas that can be used to evaluate the quality of the predicates must be determined.

Measuring the quality of the discovered patterns is an active and important area of data mining research. For example, measures for association rules such as confidence, lift and certainty factor have been used extensively (Guillet & Hamilton, 2007; Chandrasekar, Sana & Zaid, 2013). However, these measures are defined in terms of the antecedent and consequence (i.e., the structure of the model). Some of the measures that have been proposed for association rules, such as “support,” may be used as a basic inspiration for measuring the quality of fuzzy predicates; in this paper, we focus on the limitations of some of the association rule measures and how they can be adapted to fuzzy predicates. Then, we propose several new quality measures for fuzzy predicates under a different knowledge representation model.

Section 2 addresses the basic definitions of the association rules and support measure. The second important pillar in this paper is the explanation of the primary concepts of fuzzy predicates in CNF and DNF (section 3). In section 4, we present the proposed quality measures for fuzzy predicates in CNF and DNF. Section 5 shows and discusses the results that are obtained using the proposed measures with real-world datasets. Section 6 presents some concluding remarks.

## Association rules

An association rule is an expression of the form  $A \rightarrow B$ , where 'A' and 'B' are different sets of attributes. This rule can be evaluated by a number of quality measures, but “support” is one of the best-known measures that is not defined in terms of the antecedent and consequence (Agrawal & Srikant, 1994).

**Support** is the percentage of transactions that contain both A and B:

$$supp(A \rightarrow B) = \frac{|A \cap B|}{|X|} \tag{1}$$

$$|A \cap B| = \sum_{x \in X} conj(A(x), B(x)) \tag{2}$$

In (1), the numerator is the number of transactions that contain the itemset (A and B), and X is the size (i.e., number of transactions) of the database. Its values are in the range [0, 1]. If the antecedent and consequence are not present in any transaction, then it is equal to 0. If they occur in all transactions, its value is equal to 1.

An itemset with a support greater than the minimum support threshold is called a frequent or large itemset.

The disadvantage of support arises in the rare item problem. Items that occur very infrequently in the data set are deleted, although they would still produce interesting and potentially valuable rules (Sheikh & Tanveer, 2004).

If  $T = \{t_1, t_2, \dots, t_n\}$  is the database, and  $t_i$  represents one tuple in  $T$ ,  $C = \{c_1, c_2, \dots, c_m\}$  can represent all attributes of the database. Table I shows a sample database with quantitative attributes (e.g., age and income). Thus,  $T = \{t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8\}$  and  $C = \{\text{Age, Income}\}$ . For example, if the value of Income in the fourth record is required,  $t_4[c_2]$  can be used to get the value 3000.

**Table I. Sample database**

	$c_1 = \text{Age}$	$c_2 = \text{Income}$
$t_1$	26	400
$t_2$	29	950
$t_3$	26	900
$t_4$	40	3000
$t_5$	23	2300
$t_6$	64	1500
$t_7$	35	1000
$t_8$	54	2000

The theory of fuzzy sets can certainly help data mining reach this goal. The adjective “fuzzy” seems to be very popular and frequently used in contemporary studies concerning the logical and set-theoretical foundations of mathematics. Using the fuzzy set concept, the discovered patterns are more understandable to human comprehension. Fuzzy sets manage values more efficiently than existing methods because fuzzy sets soften the effect of sharp boundaries (Zadeh, 1965; Zadeh, 1975; Fayyad, Piatetsky-Shapiro & Smyth, 1996; Duarte, 1999; Han & Kamber, 2006; Venugopal, Srinivasa & Patnaik, 2009).

We can define a set of meaningful linguistic labels represented by fuzzy sets on the domain of the quantitative attributes in Table I; these are used as a new domain (Galindo, Urrutia & Piattini, 2006).

Several fuzzy sets  $F = \{f_1, f_2, \dots, f_k\}$  may be associated by attribute; for instance, if each attribute has three fuzzy sets, FAge may equal  $\{f_1 = \text{young}, f_2 = \text{middle-age}, f_3 = \text{old}\}$ , and FIncome may equal  $\{f_4 = \text{low}, f_5 = \text{medium}, f_6 = \text{high}\}$ .

This process is called fuzzification and may be performed using many of the available membership functions; for example, triangle, trapezoidal or left and right shoulder functions are commonly used because they yield good results, and their computation is simple. Other authors have proposed the use of other shapes including Sigmoid and Gaussian shapes (Cox, 1994; Mitsuishi, Endou & Shidama, 2000; Galindo, Urrutia & Piattini, 2006). Table 2 presents what could occur if the quantitative attributes (Table 1) were replaced by fuzzy attributes.

Table 2. Sample fuzzy database

	f <sub>1</sub>	f <sub>2</sub>	f <sub>3</sub>	f <sub>4</sub>	f <sub>5</sub>	f <sub>6</sub>
t <sub>1</sub>	0.6	0.4	0	1	0	0
t <sub>2</sub>	0.4	0.6	0	0.1	0.9	0
t <sub>3</sub>	0.6	0.4	0	0.2	0.8	0
t <sub>4</sub>	0	1	0	0	0	1
t <sub>5</sub>	0.8	0.2	0	0	0.4	0.6
t <sub>6</sub>	0	0.6	0.4	0	1	0
t <sub>7</sub>	0	1	0	0	1	0
t <sub>8</sub>	0	0.4	0.6	0	1	0

The standard approach to generalizing the quality measures for fuzzy association rules (Delgado, Marín, Sánchez & Vila, 2003) is to replace the set-theoretic operations, namely the Cartesian product and cardinality, with corresponding fuzzy set-theoretic operations: negation, t-norm and t-conorm (Dubois, Hüllermeier & Prade, 2003). These definitions establish families of measures, depending of the evaluation method and the quantifier of choice (Mitsuishi, Endou & Shidama, 2000).

Table 3 illustrates the computation of support for different examples of fuzzy association rules based on the database in Table 2 using the definitions of Zadeh (1965) (i.e., minimum for conjunction and maximum for disjunction).

Table 3. Support for different sample rules

Rule	Support
young → medium	0.17
middle-age → high	0.15
old → low	0

### Fuzzy predicates in CNF and DNF

Predicates are commonly used to refer to the properties of objects by defining the set of all objects that have some property in common. In general, a predicate is a statement that may be true or false depending on the values of its variables. However, in fuzzy logic, the strict true/false valuation of the predicate is replaced by a quantity interpreted as the degree of truth (Trillas, 2009).

A fuzzy predicate may be a tree where each internal node may be a fuzzy operator (e.g., conjunction, disjunction, and negation), and each leaf is a fuzzy variable of the database (Trillas, 2009). Each fuzzy variable can be associated with adverbs called hedges, which are terms that modify the shape of fuzzy sets. Hedges have two primary behaviors: reinforcement, such as "very", or weakening, such as "little" (Bouchon-Meunier & Yao, 1992).

A formula is in conjunctive normal form (CNF) if it is a conjunction of clauses, where a clause is a disjunction of literals; otherwise put, it is an AND of ORs. A formula is in disjunctive normal form

(DNF) if it is a disjunction of clauses, where a clause is a conjunction of literals. As in the DNF, the only propositional connectives that a formula in CNF can contain are AND, OR, and NOT. The NOT operator can only be used as part of a literal, which indicates that it can only precede a propositional variable or a predicate symbol.

We believe that fuzzy predicates in CNF and DNF have some grade of generality because they yield patterns that classic methods cannot obtain; also, equivalent patterns can be generated. This transformation is based on the rules of logical equivalences (Bruno, 1998) (Trillas, 2009). Even when these equivalences are more true than false in multivalued logic, fuzzy predicates in CNF and DNF is a good pattern representation to generalize knowledge.

These predicates are sometimes created by a human expert or, in the best circumstances, by the mutual consent of a group of them. However, they can also be created by algorithms that "learn" when "processing" real data. Predicate mining is a task that can be examined as an optimization problem and metaheuristics can be used to solve it.

Each fuzzy predicate in CNF or DNF can code by a vector that represents the attributes in different clauses and values. You can use positional integer encoding, where each value has a translation according to the following scale (see Table 4). In the predicate, variables can appear more than once (i.e., they may be included in two or more clauses).

Table 4. Scale

Value	Code	Meaning	Formula	Result
	0	absence	-	-
	1	F	F	0.6
F=0.6	2	not F	I - F	0.4
	3	very F	F <sup>2</sup>	0.36
	4	extremely F	F <sup>3</sup>	0.21
	5	slightly F	F <sup>1/2</sup>	0.77

An example of one predicate and its corresponding code solution is shown in Table 5.

Table 5. Predicate example

Predicate	(middle-age AND NOT high) OR (very medium)					
	middle-age AND NOT high			very medium		
Variables	f <sub>1</sub>	f <sub>2</sub>	f <sub>3</sub>	f <sub>4</sub>	f <sub>5</sub>	f <sub>6</sub>
Code	0	1	0	0	2	0
Formula	-	f <sub>2</sub>	-	-	-   - f <sub>6</sub>	-

For each predicate, the unique quality measure that is known is the FPTV (Ceruto, Rosete & Espin, 2010), which depends on the number of clauses (Z), variables (Y) and records (X) of the data set. The fuzzy value of the FPTV is in range [0, 1]:

$$TV(\wedge \text{ clause}) = \text{conj}(TV(\text{var})_1, \dots, TV(\text{var})_Y) \tag{3}$$

$$TV(\vee \text{ clause}) = \text{disj}(TV(\text{var})_1, \dots, TV(\text{var})_Y) \tag{4}$$

$$TVC = \text{conj}((TV(\vee \text{ clause})_1, \dots, (TV(\vee \text{ clause})_Z)) \tag{5}$$

$$TVD = \text{disj}((TV(\wedge \text{ clause})_1, \dots, (TV(\wedge \text{ clause})_Z)) \tag{6}$$

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

The procedure to compute FPTV is summarized in the next pseudocode.

**BEGIN**

For each record in the database (X)  
 For each clause in the predicate (Z)  
 For each attribute of the clause (Y)  
 Calculate the real value (TV<sub>var</sub>) of all attributes involved in the predicate (use Table 4)  
  
 Calculate the TV of the clauses (TV<sub>∧clause</sub> or TV<sub>∨clause</sub>) depending on the operator (equation 3-4)  
 Calculate the TV of the predicate in the record (TVC or TVF) according to the normal form (equation 5-6).  
 Calculate the final value using the universal quantifier (conjunction) in all rows (equation 7)

**END**

The FPTV is computed using fuzzy logic operators. Fuzzy logic does not give a unique definition of the classic operations such as union or intersection. Different operators that can be used imply differences in the truth value that will be obtained. Zadeh operators (i.e., min-max) are insensitive (Zadeh, 1965). In this case, the change of one argument may not change the value of the result (e.g., 0.5 ∧ 0.5 = 0.5; 0.5 ∧ 0.8 = 0.5). Probability operators (Mizumoto, 1981a) are not idempotent; the conjunction of two variables with the same values does not result in the same number (0.5 ∧ 0.5 = 0.25). Compensatory fuzzy logic is sensitive and idempotent (Espin, Fernandez, Mazcorro, Marx-Gómez & Lecich, 2006) because associativity is excluded; examples of this include the Geometric Mean and their dual (Mizumoto, 1989).

To illustrate the computation of FPTV, we show a simple example based on the database in Table 2, the predicate in Table 5 (i.e., DNF with two clauses) and the Zadeh fuzzy operator (i.e., minimum and maximum) using the pseudocode described above. In Table 6, the first three columns represent the TV of all variables. The fourth denotes the result of computing the TV of the first clause (i.e., middle-age AND NOT high) using a minimum to create the conjunction between the first and second column. The TV (clause 2) has the same value of TV (very medium).

**Table 6. Evaluation step by step**

TV (f <sub>2</sub> )	TV (f <sub>6</sub> )	TV (f <sub>5</sub> )	TV (∧ clause 1)	TVD (clause 1 ∨ clause 2)	∀x TVD(x)
0.4	1	0	0.4	0.4	
0.6	1	0.81	0.6	0.81	
0.4	1	0.64	0.4	0.64	
1	0	0	0	0	0
0.2	0.4	0.16	0.2	0.2	
0.6	1	1	0.6	1	
1	1	1	1	1	
0.4	1	1	0.4	1	

As shown in this example, the FPTV of this predicate is 0 (i.e., conjunction of the values obtained in TVD) because there is a zero in the fourth row. Veto criteria introduce this hard restriction when certain goals are not satisfied (Espin, Fernandez, Mazcorro, Marx-Gómez & Lecich, 2006). This record is only an outlier; in general, the predicate functions well in this database. If it happens during the search, the predicate obtained would be wrongly discarded.

Conversely, it always uses the universal quantifier (i.e., conjunction of all results). This operator tends to restrict the output that will

be determined by the type of fuzzy logic operator that is selected. FPTV will only return a high true value when all records have higher values.

For all of the reasons mentioned before, it is important to have others measures to evaluate the quality of fuzzy predicates in normal form to guide the search.

**New quality measures of fuzzy predicates in CNF or DNF**

The first measure proposed is the **Fuzzy Predicate Support (FPS)**. In this case, the numerator is the sum of the truth values of the predicate in each tuple without applying the universal quantifier in each transaction divided by the absolute number of transactions. FPS can be interpreted as the average support of the predicate:

$$FPS = \begin{cases} \frac{\sum_{x \in X} TVC}{|X|} & \text{If the predicate is in CNF} \\ \frac{\sum_{x \in X} TVD}{|X|} & \text{If the predicate is in DNF} \end{cases} \quad (8)$$

The second measure is the **Fuzzy Predicate Binary Support (FPBS)**. FPBS allows the determination of which percentage of records in the databases has a truth value below the threshold. Depending of the threshold selected, if the value of FPBS is low, then the predicate is not good:

$$FPBS = \frac{\sum_{x \in X} FPBTV}{|X|} \quad (9)$$

In this case, a threshold is used to calculate the **Fuzzy Predicate Binary Truth Value (FPBTV)** in each record. FPBTV can be considered *alpha-cut* (i.e., a subset of elements with membership grades of at least *alpha*).

Table 7 shows the computation of FPS and FPBS following the previous example. The selected threshold was 0.4, and all values greater than or equal to the threshold were set to 1.

**Table 7. computation of FPS and FPBS**

TVD	FPBTV	FPS	FPBS
0.4	1		
0.81	1		
0.64	1		
0	0	0.55	0.75
0.2	0		
1	1		
1	1		
1	1		

The other proposals are associated with measures of central tendency. The mean or average is affected by the asymmetry of the data distribution and the presence of "outliers." For these reasons, average pruning, a technique in machine learning that reduces the size of the instances by removing sections that may be based on noisy or mistaken data, is selected for use (Han & Kamber, 2006). To calculate it, the data are first sorted in ascending order, and then a certain percentage of data in each end of the distribution is removed.

We propose three new measures (see Table 8):

- **Fuzzy Predicate Central Pruning Average (FPCPA)**: remove 25% of the low extremes and 25% of the high extremes, and create the subset P:

$$FPCPA = \frac{\sum_{x \in P} TV(x)}{|P|} \quad P \subset X \quad (10)$$

This measure may be interpreted as the average truth value of the transaction with median truth values; it yields a version of the Fuzzy Predicate Support (FPS) that is less sensible to the extreme truth values in the transactions.

- **Fuzzy Predicate Low Pruning Average (FPLPA):** remove 50% of the lower values and create the subset L:

$$FPLPA = \frac{\sum_{x \in L} TV(x)}{|L|} \quad L \subset X, \quad L \neq P \quad (11)$$

This measure may be interpreted as the average truth value of the transaction with high truth values; it yields an optimistic version of the FPS.

- **Fuzzy Predicate High Average Pruning (FPHPA):** remove 50% of the higher values and create the subset H:

$$FPHPA = \frac{\sum_{x \in H} TV(x)}{|H|} \quad H \subset X, \quad H \neq L \quad (12)$$

This measure may be interpreted as the average truth value of the transaction with low truth values; it yields a pessimistic version of the FPS.

**Table 8. Computation of average measures**

Ranking X	Subset P	Subset L	Subset H
0			0
0.2			0.2
0.4	0.4		0.4
0.64	0.64		0.64
0.81	0.81	0.81	
	FPCPA=0,71	FPLPA=0,95	FPHPA =0,31

In addition, we propose a measure of comprehensibility, which attempts to quantify how easy it can be to understand the predicate. The generated predicates may have a large number of attributes, making them difficult to understand. In the cases of association rules, there have been some measures to manage this subject by evaluating the number of variables that are included in the rule (Martín, Rosete, Alcalá-Fdez & Herrera, 2013). The same idea may be applied to fuzzy predicates.

The **Fuzzy Predicate Comprehensibility (FPC)** is defined as:

$$FPC = \frac{1}{Y} \quad (13)$$

In equation 13, Y is the number of variables involved in the predicate. This measure may be interpreted as the inverse of the number of variables that are used in the predicate. As shown, this measure does not depend on the number of fuzzy variables in the databases. The value of FPC in the previous example is 0.33.

The fuzzy value of all measures proposed in this section is in the range [0, 1], where the value 1 indicates the best possible value (i.e., the given knowledge is true in all database); if the value drops to 0, the fuzzy value is more false than true.

The last measure proposed is the **Quantity of Zeros (QZ)** in the databases. This measure helps to determine if the value of FTPV is accurate. If the value of QZ is low, while the other measures have a high true value, then the value in question is likely an outlier, and the predicate then provides relevant knowledge.

### Experiments

This section illustrates how the proposed measures can be used to evaluate the quality of fuzzy predicates in normal form. Experiments were conducted with real-world datasets available in UC Irvine Machine Learning Repository (<http://archive.ics.uci.edu/ml/>). The algorithm used in this experiment to discover fuzzy predicates in CNF or DNF is called FuzzyPred, which was proposed by Ceruto, Rosete & Espin (2010); Rosete, Ceruto, Espin & Marx-Gómez (2011); and Ceruto, Lapeira, Rosete & Espin (2013).

Table 9 summarizes the primary characteristics of the datasets, which uses the following labels: **D** = databases; **A** = total number of attributes; **LL** = linguistic labels used; **F** = parameters of fuzzification; **R** = quantity of records. We extracted three quantitative attributes randomly from each database. The membership functions of each attribute were defined primarily by a uniform partition with trapezoidal membership functions. The linguistic label for each variable selected was also selected randomly.

We perform 30 runs, each with a maximum of 500 iterations, using several metaheuristics to obtain predicates with high truth values (FPTV) using FuzzyPred. The FPTV was computed using the Zadeh Operator (Min-Max).

**Table 10. Characteristics of the datasets**

	D	A	LL	F	R
Balance Scale (BS)	4	LW.Little	1,4 Sigmoidal	625	
		LD.Little	1,4 Sigmoidal		
		RW.High	3,5 Sigmoidal		
Basketball (BA)	5	AsisstMi.Low	.04,.08,.12,.15	96	
		Height.Medium	.12,.15,.23,.27		
		PointsMin.Medium	.23,.27,.30,.34		
Quake (Q)	4	FocalDepth.Little	-82,0,0,164,246	2178	
		Lattitude.Little	-95,-66,-8,20		
		Longitude.High	-35,36,180,251		
Pollution (P)	16	Prec.Little	3,10,22,28	60	
		Jan.Medium	3, 5, 8, 11		
		Educ.Little	-0.5,0.9,3.7,5.1		
Solar Flare (SF)	11	Activity.Little	0.8,1.0,1.4,1.6	1066	
		Evolution.High	1.8,2.2,3.0,3.40		
		HistComplex.Little	1.4,1.6,2.0,2.2		
Bolts (BO)	8	Run.Little	1.0,9.8,16.6,24.4	40	
		Speed I.High	3.6,4.4,5.3,6.0		
		Total.High	18,22,26,30		

Table 10 contains the evaluations of the best three predicates in each database using all proposed measures. The first column (**Fuzzy Predicated Identifier, FPId**) corresponds to an identifier associated with each predicate in each database. The first part of the FPId identifies the corresponding database; for example, BS<sub>2</sub> is a predicate obtained from the database Balance Scale (BS). The selected threshold to compute the measure FPBS was 0.2.

Because the number of potentially applicable predicates may be large, we illustrate one example for each database:

- **BS<sub>1</sub>:** (slightly LD.Little) AND (slightly LW.Little OR LD.Little)

- **BA<sub>1</sub>**: slightly AsisstMi.Low OR NOT Height.Medium AND (NOT AsisstMi.Low)
- **Q<sub>1</sub>**: (FocalDepth.Little) AND (Longitude.High)
- **P<sub>1</sub>**: NOT Jan.Medium AND Educ.Little
- **SF<sub>1</sub>**: (NOT Evolution.High) AND (NOT Activity.Little)
- **BO<sub>1</sub>**: (NOT Total.High) AND (NOT Run.Little)

**Table 10. Measures extended by rules**

FPIId	FPTV	FPS	FPBS	FPCPA	FPLPA	FPHPA	FPC	QZ
BS1	.70	.74	1	.75	.79	.69	.33	0
BS2	.70	.69	1	.71	.74	.65	.25	0
BS3	.5	.6	1	.61	.63	.58	.25	0
BA1	.02	.91	.97	0.95	1	0.82	.33	0
BA2	.02	.87	.97	0.93	1	0.75	.16	0
BA3	0	.59	.65	.72	.9	.18	.33	88
Q1	.36	.95	1	.97	1	.90	.5	0
Q2	.21	.94	.97	.96	1	.89	.25	0
Q3	.19	.93	.97	.96	1	.87	.16	0
P1	1	1	1	1	1	1	.25	0
P2	1	1	1	1	1	1	1	0
P3	1	1	1	1	1	1	.5	0
SF1	.5	.80	1	.87	1	.60	.25	0
SF2	0	.03	.03	.03	.06	0	.25	1028
SF3	0	.98	.99	.99	1	.97	.33	2
BO1	.33	.92	.97	.96	1	.84	.5	0
BO2	.70	.92	1	.95	1	.84	.25	0
BO3	.5	.89	1	.93	1	.78	.33	0

Analyzing the results presented in the Table 10, the following facts can be stated:

- If 50% of the highest values are removed (FPHPA), good performance is still shown; the predicate has good coverage in the database (database Q).
- When FPTV is near 1, no other measure is required for good performance because other measures would be more relaxed than those proposed and would thus not provide any new knowledge (database P).
- When FPTV=0, the veto criteria may be the cause, but another analysis is required to be certain.
- If the predicate is penalized for every measure (e.g., due to the presence of 1028 zeros at 1066 tuples, like SF<sub>2</sub>), then the quality of this predicate is poor.
- When the other measures have a high true value, then the value of interest can be considered an outlier (e.g., the value only has two zeros, like in SF<sub>3</sub>). For example, when the measure FPLPA is very small, the value of FPTV is not caused by a veto of a tuple. This conclusion can be reaffirmed by FPBS, depending on the threshold selected. This is a great example of a predicate (SF<sub>3</sub>) that can be lost if only FPTV were considered.

This section has shown how the proposed measures can be used to understand the meaning of the predicates and the real characteristics of the databases.

## Conclusions

The approach outlined in this paper justifies the use of different types of quality measures for fuzzy predicates in CNF and DNF. We compared fuzzy predicate truth values to seven other measures, three of which were statistical. The experiments show that when FPTV is near 1, no other measure is required for good performance. We suggest that the new measures are a good choice in other cases, particularly when FPTV is equal to 0 because they can help determine if the veto criteria are important. The proposed measures may be used to evaluate fuzzy predicates in different contexts independently of the way they are obtained. We also intend to use diverse and extensive test data to confirm the claims made in this paper.

## References

Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. In Proc. 20th Very Large Data Bases, 487-499.

Bruno, A. (1998). Normal forms. *Mathematics and Computers in Simulation*, 45, 413-427.

Bouchon-Meunier, B., & Yao, J. (1992.) Linguistic modifiers and imprecise categories. *International Journal of Intelligent Systems*, 7(1), 25-36.

Ceruto, T., Lapeira, O., Rosete, A., & Espín, R. (2013). Discovery of fuzzy predicates in database. *Advances in Intelligent Systems Research*, 51, 45-54.

Ceruto, T., Rosete, A., & Espín, R. (2014). Knowledge Discovery by Fuzzy Predicates. *Soft Computing for Business Intelligence*, 537, 187-196.

Chandraveer, S., Sana, A., & Zaid, M. (2013). Comparison of Interestingness Measures: Support-Confidence Framework versus Lift-rule Framework. *International Journal of Engineering Research and Applications*, 3(2), 208-215.

Cox, E. (1994). *The Fuzzy Systems Handbook: A Practitioner's Guide to Building, Using, and Maintaining Fuzzy Systems*. Boston: AP Professional.

Delgado, M., Marín, N., Sánchez, D., & Vila, M. (2003). Fuzzy association rules: general model and applications. *IEEE Transactions on Fuzzy Systems*, 11(2), 214-225.

Duarte, O. (1999). Sistemas de Lógica Difusa-Fundamentos. *Revista Ingeniería e Investigación*, 42, 22-30.

Dubois, D., Hüllermeier, E., & Prade, H. (2003). A note on quality measures for fuzzy association rules. In *Fuzzy Sets and Systems—IFSA 2003*, 346-353.

Espín, R., Fernandez, E., Mazcorro, G., Marx-Gómez, J., & Lecich, M. (2006). Compensatory Logic: A fuzzy normative model for decision making. *Investigación Operacional*, 27(2), 178-193.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD Process for Extracting Useful Knowledge. *Communications of the ACM*, 39, 27-34.

Galindo, J., Urrutia, A., & Piattini, M. (2006). *Fuzzy Databases: Modeling, Design and Implementation* (pp. 341). Idea Group Publishing.

Guillet, F., & Hamilton, H. (2007). Quality Measures in Data Mining. *Studies in Computational Intelligence*, 43.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques* (pp. 1-14) (2nd ed.). The Morgan Kaufmann Series in Data-Management Systems.

Martín, D., Rosete, A., Alcalá-Fdez, J., & Herrera, F. (2013). QAR-CIP-NSGA-II: A new multi-objective evolutionary algorithm to mine quantitative association rules. *Information Sciences*, 258, 1-28.

- Mitsuishi, T., Endou, N., & Shidama, Y. (2000). The concept of fuzzy set and membership function and basic properties of fuzzy set operation. *Journal of Formalized Mathematics*, 9(2), 315-356.
- Mizumoto, M. (1981). Fuzzy Sets and their Operations I. *Information and Control*, 48(1), 31-48.
- Mizumoto, M. (1989). Pictorial Representations of fuzzy connectives, Part II: cases of Compensatory operators and Self-dual operators. *Fuzzy Sets and Systems*, 32, 45-79.
- Rosete, A., Ceruto, T., Espin, R., & Marx-Gómez, J. (2011). A General Method for Knowledge Discovery Using Compensatory Fuzzy Logic and Metaheuristics. In *Gathering Knowledge Discovery, Knowledge Management and Decision Making* (pp. 240-271).
- Sheikh, L., Tanveer, B., Hamdani, M. (2004). Interesting measures for mining association rules. In *Proceedings of INMIC 2004* (pp. 641-644).
- Trillas, E. (2009). On a model for the meaning of predicates. A naive approach to the genesis of fuzzy sets. *Studies in Fuzziness and Soft Computing*, 243(9), 175-205.
- Venugopal, K., Srinivasa, K., & Patnaik, L. (2009). Soft Computing for Data Mining Applications. *Studies in Computational Intelligence*, 190, 1-15.
- Zadeh, L. (1965). Fuzzy Sets. *Information Control*, 8, 338-353.