# APPLYING FUZZY COMPARATORS ON DATA MINING

Angélica Urrutia[1], José Galindo[2], Leonid Tineo[3], José Morales[1] and Claudio Gutiérrez[4]
[1] *Universidad Católica del Maule, Talca, Chile*
[2] *Universidad de Málaga, Málaga, Spain*
[3] *Universidad Simón Bolívar, Caracas, Venezuela*
[4] *Universidad del Bío-Bío, Concepción, Chile*

Keywords: Fuzzy comparators, Data mining, Complex queries with data mining algorithms.

Abstract: This paper presents a fuzzy comparators module for Data Mining. This module allows querying data obtained by the application of existing data mining algorithms in SQL Server 2008. It provides the end user of a tool that gives useful information and knowledge about variables that have direct impact on the analysis of management indicators. The main contributions of this work are: first analysis and implementation of algorithms relaxed using fuzzy comparators, second deployment in a case to analyze the results.



## 1 INTRODUCTION

Our goal is to provide a better information service to the end user requiring data analysis and management indicators. We propose the use of possibility and necessity fuzzy comparators in Data Mining based queries. Such comparators are part of fuzzy logic and have their foundations in fuzzy sets theory. In previous works, this theory has been used to extend the Relational Data Model and its implementations, thus emerging the Fuzzy Relational Databases (Galindo, 2008). One remarkable effort in this sense is FSQL, an extension of SQL that provides imprecise data processing and vague queries capability (Galindo et al., 2006). In such work the application domain was in the area of tourism (Carrasco et al., 2008). That previous work extended FSQL use in mining query, giving thus more interesting results of mining process. Other extensions of fuzzy sets applied to Data Mining study cases can be found in (Urrutia et al., 2010). In this work, it is possible to provide to the user the ability to consult with possibility and necessity fuzzy comparators, extending the Data Mining Methods and applied to analysis of data obtained with Cluster technique in Data Mining (Fiel and Abonyi, 2008). Decision Trees, Naïve Bayes and Clustering, which are included in SQL Server 2008 (Urrutia et al., 2010).

Our proposal foundation is in fuzzy sets theory (Zadeh, 1865). On an universe $U$, a fuzzy set $A$ is defined by a *membership function* $\mu_A$ and $A = (\mu_A(u) / u: u \in U, \mu_A(u) \in [0,1])$. Where, $\mu_A(u)$ is *the membership degree* of the element in $A$. Here $\mu_A(u)=0$, indicates that $u$ does not belong at all to the fuzzy set $A$ *and* $\mu_A(u) =1$ indicates that $u$ belongs entirely to $A$. A data or imprecise value can be represented by the fuzzy set of possible values. This is what is called a possibility distribution. These concepts can be used in order to define fuzzy comparators, as proposed in previous works (Urrutia et al., 2008). In Possibility Theory there are two measures of truth: the possibility and necessity measures. This leads to distinguish between necessity and possibility fuzzy comparators.

Proposal using comparators for data crisp or label in Data Mining result. In classical Data Mining algorithms input data and results are usually crisp attributes and values. In order to implement this type of tool, it is necessary to extend other datatypes and more flexible queries, therefore, it is possible to obtain a better analysis of information provided by the algorithms. In a study conducted in our group of databases, we generated two types of extensions to the query results applied to a Data Mining algorithm. First using data comparators and crisp. Second using data fuzzy comparators use of necessity and possibility tagged data.

This paper is organized as follows. Section 2 includes a background about fuzzy comparators. Sections 3 and 4, we propose a layer for data mining. In Section 5, the experimental environment


Urrutia A., Galindo J., Tineo L., Morales J. and Gutiérrez C..
APPLYING FUZZY COMPARATORS ON DATA MINING.
DOI: 10.5220/0003670704820485
In *Proceedings of the International Conference on Evolutionary Computation Theory and Applications* (FCTA-2011), pages 482-485
ISBN: 978-989-8425-83-6

is documented and the results are analyzed.

## 2 BACKGROUND ON FUZZY COMPARATORS

Classical comparison operators are "equal" (=), "greater than" (>), "less than" (<), "greater than or equal to" (≥), "less than or equal to" (≤) and "not equal" (≠). These comparators are used to compare numbers, texts and dates. Fuzzy comparators are an extension of classical comparators, very useful for fuzzy databases with fuzzy queries (Zadrozny et al., 2008). The GEFRED model defines a type of general comparator based on existing classical comparators. The only requirement is that fuzzy comparators should respect the results of classical comparators when comparing crisp data. This theoretical base was used by FSQL (Galindo et al., 2008) to define a complete family of fuzzy comparators (see Table 1).

Table 1: Fuzzy Comparators of FSQL.

| Fuzzy Comparator | Meaning |
|---|---|
| FEQ, NFEQ | Possibly Equal, Necessarily Equal |
| FDIF, NFDIF | Possibly Fuzzy Different to, Necessarily Fuzzy Different to |
| FGT, NFGT | Possibly Greater Than, Necessarily Greater Than |
| FGEQ, NFGEQ | Possibly Greater or Equal, Necessarily Greater or Equal |
| FLT, NFLT | Possibly Less Than, Necessarily Less Than |
| FLEQ, NFLEQ | Possibly Less or Equal, Necessarily Less or Equal |
| MGT, NMGT | Possibly Much Greater Than, Necessarily Much Greater Than |
| MLT, NMLT | Possibly Much Less Than, Necessarily Much Less Than |

FSQL allows fuzzy comparators on unordered underlying domains (of course, only `FEQ`, `NFEQ`, `INCL`, and `FINCL`) for details.

Necessities comparators are more restrictive than possibility ones, i.e. their fulfillment degree is always lower than the fulfillment degree of their corresponding possibility comparator. Note that possibility comparators measure how possible it is that the condition is satisfied, whereas necessity comparators requires that the condition is satisfied in some degree. Thus, necessity comparators do not satisfy the reflexive property.

On the other hand, there are comparators whose results include others. For example, in crisp mode, the result of the comparator >= includes the result of >. We can then say that the comparator > is more restrictive than >=. This means that more restrictive comparators will select a smaller or equal number of tuples, and these selected tuples will never have a greater fulfillment degree than with less restrictive comparators.

## 3 PROPOSED LAYERS 1 AND 2 IN OUR APPLICATION

In order to validate our proposal, we used a case study. The following describes the work done by each layer proposal.

*Layer 1:* Our study database is Adventure Works Cycles included in SQL Server 2008. In defining the problem we used a partial data model data warehouse. This was the input to the different implementations of the Data Mining process. The scenario selected was the Direct Mail, and three algorithms were implemented: Decision Trees, Clusters and Naive Bayes.

The indicator for this scenario is the best answer from user's point of view, i.e., how likely is that a person with certain characteristics buys any offered product. To be more specific: Which is the probability that each customer buy a bicycle?. We must analyze which of these three algorithms work better. For this there is a tool called "Lift Chart" which can be found under the tab "Data Mining Accuracy Chart" from SQL Server.

*Layer 2:* For the Direct Mail Scenario and the indicator as defined above, we apply the algorithm that best fits our ideal model (perfect prediction). In this case, we have chosen the Decision Tree algorithm. We will make predictions with this algorithm and, subsequent, classical and fuzzy queries.

When using the Decision Trees algorithm to make predictions, it generates a prediction query on a table of cases. This query computes the probability that - every person in the case table buy a product-. This table of cases contains profiles of likely customers. It stores the probability that each potential buyer purchases a product (in this case a bicycle).

## 4 EXPERIMENTAL ENVIRONMENT: LAYER 3

As a result of the *Layer 2*, the output data through

the Decision Trees algorithm are showed. *Layer 3* consists on extending the classical queries to fuzzy queries using fuzzy comparators. For each one of the eight possibility fuzzy comparators (see Table 1) considered in this work, nine queries were applied over the results obtained by the Data Mining. Despite the fact necessity comparators should be used, we only used the possibility fuzzy comparators as a simple way of showing the test results.

Three types of queries were considered: Classical query, fuzzy queries using approximately values and other fuzzy queries using other more generic linguistic labels where the value to take are low, medium and high.

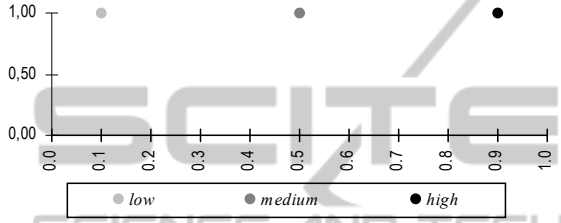


Figure 1: Standards for precise values.

The classical query takes the following values: 0.1, 0.5 and 0.9 (values for low, medium and high respectively as Figure 1 shows). The precise comparator is = instead of FEQ, < instead of FLT, <= instead of FLEQ, > instead of FGT, >= instead of FGEQ, Fuzzy comparators MLT and MGT have no similar operators in SQL classic comparators.

In the fuzzy queries, attribute *numeroDifuso* is compared with three approximate values: “approximately 0.1” (low level), “approximately 0.5” (medium), and “approximately 0.9” (high) see Figure 2.

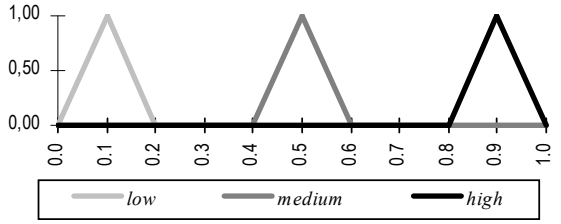


Figure 2: Standards for fuzzy values.

Finally, the fuzzy queries may also be done using linguistic labels. We have defined three labels: low, medium, and high. The definition of these labels is depicted on Figure 3.

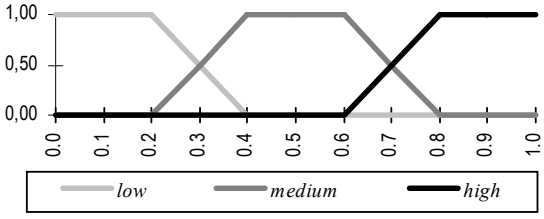


Figure 3: Standards for linguistics level values.

In order to show the extensions in the analysis of results of data mining using fuzzy comparators, one experiment based in a study case had been defined. This experiment has three query classifications: accurate queries, approximate queries, and linguistic queries. The obtained results as well as the SQL statements are showed below.

In the experimental performance, three variables were observed: The number of resulting rows, the execution total time of query, and frequency distribution on the satisfaction grade of the rows in the result set.

# 5 ANALYSIS THE RESULTED

In (Urutia et al., 2010), several qualities of fuzzy queries are showed. Generally, the use of fuzzy comparators in any query increases the number of selected tuples. Thus, it is possible to retrieve a large amount of possible answers. Moreover, the comparators FEQ/NFEQ, FLEQ/NFLEQ and FGEQ/NFGEQ give more results with linguistic labels than with approximate numbers. The mentioned results are given by the increase of value ranges in core of the set that defines to the fuzzy number. Nevertheless, the comparators FLT/NFLT, FGT/NFGT, MGT/NMGT and MLT/NMLT give more results when they use an approximate number than when they use a linguistic label.

The results obtained in this experiment can be classified by the number of obtained tuples, which give the next analysis:

- If the results are better and more interesting.
- The results appear ordered by their “fuzzy fulfillment degree”. It allows to obtain a set of ranked rows and then the user may use the best rows (or even the worst ones).
- It is possible to admit data in order to process of fuzzy manner.
- If we want less rows, in fuzzy queries we can use necessity comparators, we can increase the fulfillment threshold in each fuzzy condition, or we can modify the fuzzy constants in the conditions.

The frequency distributions about the satisfaction degree were analyzed. Nevertheless, it is impossible to show here because it provides about seventy three graphs. In the cases of precise queries, all results have satisfaction degree 1. In summary, in the queries with fuzzy quantifiers and approximate number, varied CDEG values tended to be low were obtained. The queries with linguistics labels, the degree were more distributed because it was possible to obtain high values as well as low values. These results show that the precise queries do not help user in order to discriminate answers. On the other hand, the use of fuzzy comparators give different satisfaction grade on the answers, therefore, these can help to the final user.

## 6 CONCLUDING REMARKS

Fuzzy comparators used on data, which have been obtained from a Data Mining application, are a very powerful tool when working with indicators to support management decision making.

Fuzzy comparators may deliver more rows, depending on the thresholds, and the used fuzzyfiers, for analysis of data obtained by a data mining algorithm, which can support better decision making. Furthermore, the rows may be retrieved ordered by the fulfillment degree. The fuzzy values for comparison are important to be well defined, but such definition should be done together with the end user, because the queries results must be in correspondence with management indicators. We are currently working to incorporate linguistic labels for querying data mining results. It will give another spectrum analysis queries data mining algorithms. Other possible future contribution would be giving importance degrees to the data from each cluster of Data Mining and also extend to fuzzy association rules, or degrees of membership of data to the cluster. As future work, it is proposed to extend other components of fuzzy logic to data or Data Mining algorithms (Feil and Abonyi, 2008). We also analyze different types of query terms fuzzification.

## REFERENCES

Carrasco R., Araque F., Salguero A., Vila M. A., 2008, Applying Fuzzy Data mining to Tourism Area. In *Handbook of Research on Fuzzy Information Processing in Databases*, Vol. II, pp. 563-589. Information Science Reference (http://www.info-sci-ref.com).

Galindo, J. (Ed.). 2008, Section IV Fuzzy Data Mining. *Handbook of Research on Fuzzy Information Processing in Databases*. Hershey, PA, USA: Information Science Reference (http://www.info-sci-ref.com).

Galindo, J., Urrutia, A., Piattini, M. 2006, Fuzzy Databases: Modeling, Design and Implementation. *Idea Group Publishing Hershey*, USA.

Feil, B. & Abonyi, J., 2008, Introduction to Fuzzy Data Mining Methods. In *Handbook of Research on Fuzzy Information Processing in Databases*, Vol. I, pp. 55-95. Hershey, PA, USA: Information Science Reference (http://www.info-sci-ref.com).

Urrutia A., Tineo L. & Gonzalez C., 2008, FSQL and SQLf: Towards a Standard in Fuzzy Databases. In *Handbook of Research on Fuzzy Information Processing in Databases*, Vol. I, pp. 270-298.

Urrutia A., Gutiérrez C., Méndez J., 2010, Consultas con Comparadores Difusos en Algoritmos de Minería de Datos: SQL Server 2008. *XV Congreso Español S Tecnologías y Lógica Difusa ESTLYF 2010*. Huelva España. Febrero 2008.

Zadeh L. A. 1965, Fuzzy Sets. *Information and Control*, 8, pp. 338-353.