Selection and Classification of Statistical Data Using Fuzzy Logic

Miroslav Hudec[(1)], Mirko Vujošević[(2)]
[(1)] INFOSTAT – Institute of Informatics and Statistics, Bratislava, Slovakia
[(2)] Faculty of Organizational Sciences, Jove Ilića 154, Beograd, Serbia

Abstract

Linguistic expressions like: high rate of unemployment, or high migration level etc., are very often used in life and in statistics. The goal of this research is to capture these expressions and make them suitable for database queries and classification tasks. This paper shows differences between usual and fuzzy approaches in database queries and classification and points out advantages of the fuzzy approach. The fuzzy logic deals with reasoning that is approximate rather than precise to solve problems in a way that more resembles human logic. For querying process the generalised logical condition in the WHERE part of the SQL was developed. In order to classify data, fuzzy queries are generated from fuzzy rules. The objective of integration is to use the same generalised logical condition in data selection and data classification. The proposed fuzzy approach provides flexibility when users cannot unambiguously set the boundaries between data. The fuzzy approach also extracts additional valuable hidden information in comparison with the usual, crisp approaches. In this way, queries based on linguistic expressions on client's side are supported and are accessing relational databases in the same way as the classical SQL enables.

Keywords: fuzzy SQL, fuzzy classification, database.

1. Introduction

This paper examines two often used processes: data selection (database queries) and data classification. The aim of this paper is to emphasize situations when classical {true, false} logic is not adequate in these two processes and offers fuzzy logic because the fuzzy logic is an approach to computing based on "degrees of truth" rather than the usual "true or false" logic.

Fuzzy approach is suitable for statistical databases. Linguistic expressions like: high rate of unemployment or medium migration level etc., are very often used and it is useful to catch them and use in database queries and classification. Statistical indicators are often collected with some errors and vagueness and classical techniques may involve some inadequately selected, or classified data.

This paper after short introduction of fuzzy logic idea presents our research in fuzzy database querying area, where we have created generalized logical condition (GLC). This fuzzy query idea is explained in one case study. During research in database queries and classification by fuzzy systems we got a new idea of data classification technique. The new approach to treating classification is based on the same GLC. This

idea leads to integration of data selection and classification into one tool. In this paper the characteristics of a new approach are explained on one case study. The idea how to improve the whole solution to satisfy all Boolean axioms is mentioned and finally some conclusions are drawn.

2. Fuzzy idea

The core of both classical and fuzzy logic is the idea of a set. In classical set theory an element belongs or does not belong to a set. For example consider a set called high unemployment (HU) defined as follows: HU={x| unemployment(x)>=10%} where x is a region. It means that region with 9.95% unemployment does not belong to the HU but region with 10% belongs. These constraints are drawback when the boundaries between values of some attributes are continuous.

The fuzzy logic theory brings a paradigm in work with the graduation, uncertainty and ambiguity described by linguistic expressions. The fuzzy set theory permits the gradation of the membership of the element in a set. This gradation is described by a membership function μ valued in the interval [0, 1]. The HU example can be presented by fuzzy sets shown in figure 1. User could define that the unemployment equal and bigger than 10% (L_p in figure 1) is HU, the unemployment smaller than 8% (L_d in figure 1) definitely is not HU and unemployment between 8% and 10% partially belongs to the HU concept. The fuzzy approach uses knowledge that does not have clearly defined boundaries. Many of the phenomena from real world fall into this class.

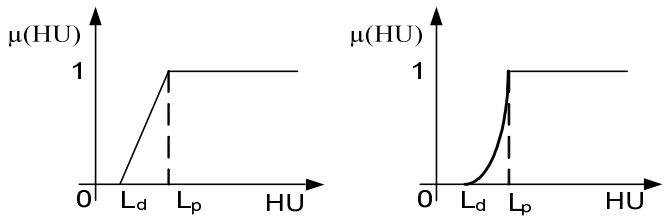


Figure 1: Fuzzy sets for big unemployment concept

3. Database queries

An answer for the question why users search databases could be as follows: To retrieve data needed to make a decision. The structured query language (SQL) is used to obtain data from relational databases. The SQL query is as follows:

```
select attribute_1,…,attribute_n
from T                                                    (1)
where attribute_p > P and attribute_r < R.
```

The result of the query is shown in graphical mode in figure 2. Values P and R delimit the space of interesting data. Small squares in the graph show database records. In the graph is obviously shown that two records are very close to satisfying the query criterion. These two records could be two potential customers and direct marketing could attract them or two municipalities which almost satisfy criterion for some financial support.

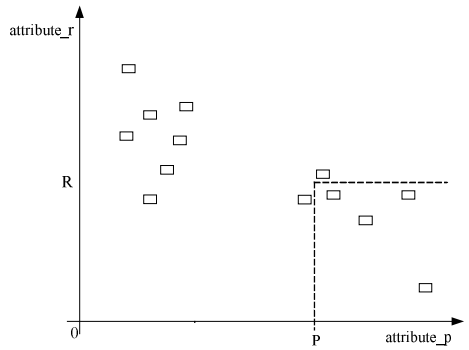


Figure 2: The result of a SQL query

SQL makes crisp selection. It means that the record would not be selected even if it is extremely close to the intent of a query. This is the penalty paid for using the crisp logic in selection criterion. In cases when the user can not unambiguously separate interested data from not interested by sharp boundaries or when the user wants to obtain data that are very close to meet the query criterion and to know the index of distance to full query satisfaction, it is necessary to adapt the SQL to these requirements.

3.1 Fuzzy SQL queries

A fuzzy query system is an interface to users to get information from database using (quasi) natural language sentences. The answer to a fuzzy query sentence is generally a list of records, ranked by the degree of matching (Branco et. al., 2005).

The query compatibility index (QCI) is used to indicate how the selected record satisfies a query. The QCI has values from the [0,1] interval: 0 – record does not satisfy a query, 1-record has full query satisfaction, interval (0,1) – record partially satisfies a query.

The GLC formula for the WHERE part of the SQL based on linguistic expressions was built in (Hudec, 2007). The implementation of GLC for statistical databases is explained in (Hudec, 2008). In this paper the process of fuzzy selection by GLC is mentioned and examined in a case study.

The GLC has the following structure:

$$\text{WHERE} \overset{n}{\underset{i=1}{\otimes}} (a_i \, \mathbf{o} \, L_{ix}) \tag{2}$$

where n denotes number of fuzzy constraints in a WHERE clause of a query,

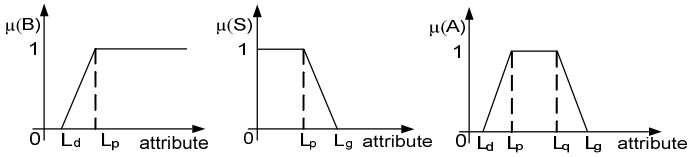
$$\otimes = \begin{cases} and \\ or \end{cases}$$

where *and* and *or* are fuzzy logical operators,
and

$$a_i \, \mathbf{o} \, L_{ix} = \begin{cases} a_i > L_{id}, & \text{a}_\text{i} \text{ is Big} \\ a_i < L_{ig}, & \text{a}_\text{i} \text{ is Small} \cdot \\ a_i > L_{id} \text{ and } a_i < L_{ig}, & a_i \text{ is About} \end{cases}$$

where a_i is a database attribute, L_d is the lower bound and L_g is upper bound of a linguistic expression described by fuzzy set shown in figure 3 .

The querying process consists of the two steps. In the first step lower and/or upper bounds of linguistic expressions (fuzzy sets) are used as parameters for database queries. It means that all records that have QCI greater than zero are selected. In the second step the chosen analytical form of the fuzzy set is used to calculate the membership degree of each selected record to appropriate fuzzy set. Finally, appropriate t-norms or t-conorms are used to calculate QCI values for all retrieved records.



a) Big value b) Small value c) About value

Figure 3: Fuzzy sets

In many-valued logic there exist many functions describing *and* operator (t-norms) and *or* operator (t-conorms) because each of conditions inside the WHERE clause can be partially satisfied. The following functions can be used for t-norms:

- minimum: $QCI = \min(\mu_i(a_i)), \quad i=1,...,n$ (3)
- product: $QCI = \prod (\mu_i(a_i)), \; i=1,...,n$ (4)
- bounded difference (BD): $QCI = \max(0, \sum_{i=1}^{n} \mu_i(a_i) - n + 1)$ (5)

where $\mu_i(a_i)$ denotes the membership degree of the attribute a_i to the i-th fuzzy set. Min t-norm takes into account the lowest value of membership degrees to fuzzy sets. Product t-norm takes into account all membership degrees in the WHERE clause.

A fuzzy query interpreter which transforms fuzzy queries to the SQL structure was developed. In this way, queries based on linguistic expressions on the client side are supported and accessing relational databases in the same way as the SQL is enabled.

3.2 Case study

This system is tested on data from the Urban and Municipality Statistics database used in the Statistical Office of the Slovak Republic (Benčič and Hudec, 2002). In this case study, districts with high length of road and small area size are sought. The big road infrastructure density is analyzed as an illustrative example. The query has the following form:

```
select district, roads,  area
from T
where roads is Big and area is Small
```

The length of road indicator is represented by „Big value“ fuzzy set with these parameters L_d=200 km and L_p =300 km and the shape as from Figure 3a). The „Small value“ fuzzy set with parameters L_p=450 km^2 and L_g =650 km^2 and shape as from Figure 3b) describes the area attribute

Result of fuzzy query is shown in Table 1. The value of min t-norm (3) is used for district ranking. The Table 1 shows six districts fully satisfying the query; one district is extremely close to satisfying the query and another two districts are close to the

query criterion. It means for example that even small changes in districts attributes could involve that another records fully satisfy the query. If SQL was used, this additional information would remain hidden.

Table 1: Result of fuzzy query.

District	Roads [km]	Area [km^2]	QCI
Detva	567,2	449	1
Myjava	563,9	327	1
Žarnovica	366,6	426	1
Bratislava I	335,1	10	1
Púchov	320,9	375	1
Piešťany	305,6	381	1
Považská Bystrica	324,5	463	0,935
Kysucké N. M.	269,9	174	0,7
Senec	269,1	360	0,69
Žiar nad Hronom	249,8	518	0,5
Nové Mesto n. V.	528,5	580	0,35
Krupina	334,9	585	0,325
….	….	…..	…..

In cases when user uses SQL and wants to obtain similar results like result presented in Table 1 it is needed to make small changes in criterion parameters and to execute larger number of queries. The query from example (1) could be modified as follows:

- **(attribute_p > P and attribute_r < R)** extract records that satisfy initial conditions.
- **(attribute_p > P-p_1 and attribute_p < P) and (attribute_r < R+r_1 and attribute_r < R)** select records that meet query criteria with value of 0.9 or almost meet the query criteria (where p_1 and r_1 are small real values greater than zero),
- **(attribute_p > P-p_2 and attribute_p < P-p_1) and (attribute_r < R+r_2 and attribute_r < R+r_1)** select records that meet query criteria with value of 0.8 or records that are very close to query criteria (where $p_1<p_2$, $r_2>r_1$),
- etc.

The following conclusion appears: for the very soft gradation, the unlimited number of SQL queries has to be used. In case of fuzzy queries, only one query meets this requirement.

4. Data classification

Users classify data in order to find in which class each of classified record (territorial unit, customer) belongs. Expert systems are one of techniques capable of data classification. The usual classification by expert system is illustrated on the example of estimating and planning of road maintenance needs in the winter. Municipalities are classified according to the length of roads in km and the number of days with snowing. Classification diagram is presented in figure 4. The figure 4 shows the classical classification. Municipalities are divided into four classes from class C1 (the smallest needs for the maintenance) to class C4 (the biggest needs for the maintenance). This method treats the top rated territorial unit T4 in the same way as T3. Units T2 and T3 have similar length of roads as well as similar number of days with snowing. However, T2 and T3 are treated in different classes.

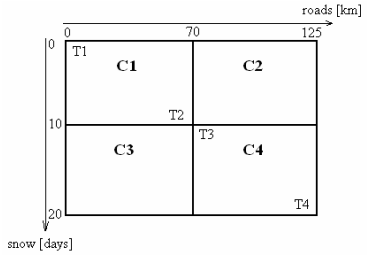


Figure 4: Classical classes

4.1 Fuzzy classification

Fuzzy classes reflect reality better and allow decision makers or analytics to describe input attributes and output classes more intuitively using linguistic variables, overlapping classes and approximate reasoning. Objects that belong to more than one class are treated in all classes where they have partial membership.

In order to solve a fuzzy classification problem within a knowledge-based fuzzy inference system (FS) it is necessary to fuzzify attributes, determine all IF-THEN rules (rule base), process them and to provide result in a usable and understandable form. More about fuzzy classification is in (Hudec and Vujošević, 2005).

The advantages of fuzzy systems are as follows. They

- enable the creation of logical inference system based on human mind including uncertainties in membership degrees to the appropriate fuzzy sets.
- support the inference process based on "IF-THEN" rules.
- enable accessible, understandable and easy to use and modify knowledge base.

There are many fuzzy system softwares capable to solve classification tasks, for example MatLab or FLOPS. These softwares have been produced to solve a wide area of tasks but they are complicated for users. In order to solve a classification task, the decision maker needs the assistance to prepare the input data from database into proper format for the FS and to present the results into a useful and understandable form. This part could be programmed but it is not a trivial task. The decision maker also needs assistance to set the most suitable mathematical functions inside the FS.

4.2 Integration of data selection and classification

If it is goal to create easy to use soft computing tool for classification, fuzzy systems are not very suitable from users point of view. The new idea for classification and how to integrate data selection and data classification has been found during our work on fuzzy queries. IF part of a fuzzy rule corresponds to a WHERE clause of a query, and all rules related to the same output class (THEN part of a rule) are aggregated with an OR operator into the same query. A fuzzy query would return all records together with their membership degrees to the query criteria satisfaction. This membership degree is also membership degree to the appropriate output class.

The classification query language is designed in the spirit of the above described fuzzy SQL. The difference is in the added clause *classify_into*. The *classify_into* clause specifies the name of the output class to which selected records are classified. The structure is as follows:

classify_into [class$_i$]
select [attribute$_1$],….[attribute$_n$]
from [tables, relations]
where $\bigoplus_{k=1}^{K} \bigotimes_{i=1}^{n} (a_i \; \mathbf{0} \; L_{ix})$

where $\otimes$ is AND operator, n is the number of attributes inside the IF part of the rule, $\oplus$ is OR operator which connects those k antecedents in IF part that have common THEN part or the same output class.

The results of all queries are objects selected into overlapping classes. The final rank for each record can be calculated from the equation:

$$R_O = \sum_{i=1}^{m} \mu_{Oci} P_i \tag{6}$$

where m is number of classes, μ_{Oci} is the membership degree of object O to class C_i and P_i is coefficient describing class C_i.

Advantages of this approach are as follows:

- Queries select only records that will be classified. Records that do not belong to any class are not needlessly selected;
- Data preparation to adequate input vector or matrix is not needed;
- Presentation of results in useful and understandable form for example in xls format or on thematic maps could be easy implemented.

4.3 Case study

The system is applied and tested for municipality classification using data of Banská Bystrica region from the same Urban and Municipality Statistics database. In this case study municipalities are classified according to the percentage of needs for the winter road maintenance. In this example two attributes are used and fuzzified into two sets: length of roads in kilometers (Road) and number of days with snow (Snow). These sets are shown in figure 5.

This example contains four fuzzy rules with the following structure:

- If Road is Small and Snow is Small Then Maintenance is Small.
- If Road is Small and Sow is Big Then Maintenance is Medium.
- If Road is Big and Snow is Small Then Maintenance is Medium.
- If Road is Big and Snow is Big Then Maintenance is Big.

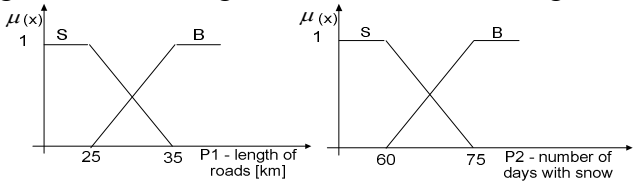


Figure 5: fuzzy sets Small (S) and Big (B) for Roads and Snow indicators.

Three fuzzy queries are created from these four rules. The query for output class Medium is as follows:

Classify_into M
select municipality, roads, snow
from T
where (roads is Small and snow is Big) or (roads is Big and snow is Small)

The percentage of needs can be associated with each output class. For instance, class S (Small) gets a percentage of needs of 10% or according to equation (6): P_s=0.1, class M (medium) gets 50% and municipalities from class B (Big) get 90% from considered needs.

This tool classified 126 municipalities. 10 of them fully belong to class S, 74 fully belong to class M, 14 fully belong to class B and other 28 partially belong to more than one class. If classical classification was used, this additional valuable information would be hidden. Table 2 shows ranking results for some municipalities.

Table 2: Some of classified municipalities.

Municipality	Coefficient of needs (P)	Municipality	Coefficient of needs (P)
Radzovce	0,1	D. Harmanec	0,5
Dudince	0,1	Kremnica	0,5
Lipovany	0,1268	Trnavá Hora	0,5
Tornaľa	0,2103	Poltár	0,5146
Kalinovo	0,2868	H. Tisovník	0,62
Hajnáčka	0,34	Donovaly	0,86
Vinica	0,34	B. Bystrica	0,9
Čebovce	0,3932	Ľubietová	0,9

From membership degree to each class and equation (6) the coefficient of needs (R) is calculated. Two municipalities are taken as example:

- **Čebovce** belongs to class S with degree of 0.267 and M with degree of 0.733. P(Čebovce)=0.267*0.1+0.733*0.5=0,3932.
- **Banská Bystrica** fully belongs to class B. P(Banská Bystrica)=1*0.9=0.9.

In case of classical classification two municipalities with very similar indicators values near the boundary value may be classified into different classes and it will cause difference between obtained needs. To avoid this disadvantage the user has to create significantly greater number of output classes and rules if he wants to use crisp tools in comparison with the fuzzy approach. Indeed for very soft ranking by crisp tools user needs nearly infinity number of input ranges, rules and output classes.

5. Discussion about proposed model

The usefulness of proposed fuzzy querying approach depends also on the theoretical and practical development of fuzzy database systems. In fuzzy databases the fuzzy query is a part of the fuzzy database management system and a special fuzzy query development is not necessary. Many statistical databases are developed in relational database management systems. This trend dominates in development of new statistical information systems and databases so the fuzzy selection presented in this work have the significant perspective for further usage. However, further development of

integrated selection and classification does not depend on fuzzy database development and its practical applications.

The GLC is developed on axioms of fuzzy logic. Fuzzy logic is based on the same principle as classical logic, the principle of truth-functionality. Logic is truth functional if the truth value of a compound sentence depends only on the truth values of the constituent atomic sentences, not on their meaning or structure (Radojević, 2008).

In case of all many-valued logics, including fuzzy logic, this principle is not suficient and as a consequence these logics are not in the frame of Boolean algebra (BA). Zadeh, the innovator of fuzzy sets and fuzzy logic, described fuzzy logic as precise many-valued logic where axioms of contradiction and excluded middle are not satisfied. First way how to use gradation in mathematics is to leave these axioms and accept the principle of truth functionality with all consequences. Radojević in (Radojević, 2008) got positive answer to the question: Can the idea of fuzziness be realized in a Boolean frame? New approach to treating fuzzines or gradation in logic is based on interpolative realization of Boolean algebra (IBA). IBA has symbolic level (finite BA) and semantic or valued level.

Above mentioned problem does not appear in fuzzy classification. There are some situations when this problem can arise in fuzzy querying. It is possible to avoid this problem by selecting adequate t-norm or t-conorm function for queries. Although this kind of choosing adequate functions satisfies demands, it is very interesting to improve fuzzy queries with IBA concept. The IBA could provide better flexibility and avoid theoretical possible situations when inappropriate functions are chosen.

The web application with a fuzzy module for data dissemination is another way of improvement of this fuzzy query approach. Statistical institutions put vast amount of data into their websites and providing a selection criteria by linguistic expression gives natural way for data selection.

6. Conclusion

It is proven in our research that the proposed fuzzy system for selection and classification may be successfully used. If crisp sets and sharp boundaries in queries are used then the small error in data values or cases when user can not unambiguously define the criteria by crisps sets may involve some inadequate selected or classified data. Fuzzy logic provides answer how to avoid these disadvantages. User gets two advantages: the model is described in natural language and the result gives additional valuable information that might remain hidden if usual methods are used.

FS are powerful tools but it is not easy to work with them. Both, advantages of FS and advantages of our querying process are included in our classification concept. The knowledge base is understandable, easy to modify and use. Our classification approach exempts users from data preparation to adequate input matrix, preparation of results into adequate form and setting appropriate functions inside FS.

The fuzzy methods are in this approach independent modules. The modularity of here mentioned modules allow their modifications and improvements independently. Implementation of the IBA in querying process is very interesting topic for further research.

The formulas for transforming fuzzy queries to the classical ones and fuzzy rules to fuzzy queries were built and a fuzzy query interpreter based on these formulas is under developing. In this way, queries based on linguistic expressions on client side are supported and accessing relational databases in the same way as the classical SQL is enabled. The interpreter also creates fuzzy queries from fuzzy rules and converts them into classical SQL. At the end of classification process interpreter puts records into appropriate classes. No modification of databases has to be undertaken.

This approach could be reused for another databases or purposes. The core of this tool (GLC) remains the same, only an input and output parts have to be adapted to achieve new needs. If users are prepared to accept a less accurate system that contains approximate reasoning, they look for one which is fully comprehensible to them. Our approach could meet their needs.

References

Benčič A., Hudec M. (2002) MOŠ/MIS–Urban and municipal statistics project and information system of the Slovak Republic, *Symposium on operational research,* Vuletić Print, XXI-32--XXI-35

Bosc P., Pivert O. (2000) SQLf Query Functionality on Top of a Regular Relational Database Management System, in: *Knowledge Management in Fuzzy Databases,* Pons M. et al. (Eds.), Physica Publisher 171-190.

Branco A., Evsukoff A., Ebecken N. (2005) Generating Fuzzy Queries from Weighted Fuzzy Classifier Rules, *ICDM workshop on Computational Intelligence in Data Mining*, IOS Press, 21-28

Cox E. (2005) *Fuzzy modeling and genetic algorithms for data mininig and exploration*, Morgan Kaufmann Publishers. San Francisco.

Hudec M., Vujošević M., 2005. Fuzzy systems and neuro-fuzzy systems for the municipalities classification. In *EUROFUSE*, *2005, Eurofuse anniversary workshop on "Fuzzy for Better"*. M. Pupin Institute, 101-110.

Hudec M. (2007) Fuzzy improvement of the SQL. *Balkan*-Conference-*on*-*Operational*-*Research,* http://balcor.fon.bg.ac.yu/unos/Docs/sec06/4.pdf.

Hudec M. (2008) Fuzzy SQL for statistical databases, *UNECE-EUROSTAT-OECD Meeting on Management of Statistical Information Systems*, http://www.unece.org/stats/documents/ece/ces/ge.50/2008/wp.12.e.pdf.

Radojević D. (2008) Interpolative Realization of Boolean Algebra as a Consistent Frame for Gradation and/or Fuzziness, in: *Forging New Frontiers: Fuzzy Pioneers II Studies in Fuzziness and Soft Computing*, Nikravesh M. et al. (Eds.), Springer, 295-318.

Siler W., Buckley J. (2005) *Fuzzy expert sytems and fuzzy reasoning*. John Wiley & Sons, Inc. New Jersey.

Werro N. et al. (2008) Concept and Implementation of a Fuzzy Classification Query Language, *International Conference on Data Mining*, CSREA Press, 208-214.