

# UNCERTAINTY SUPPORT CLASSIFICATION IN RELATIONAL DATA SETS: A FUZZY LOGIC BASED APPROACH

C. Amanatidis, M. Halkidi, M. Tzouris, M. Vazirgiannis  
Dept of Informatics,  
Athens Univ. of Economics & Business, Athens, Greece (Hellas)

## ABSTRACT

The KDD process aims at searching for interesting instances of patterns in data sets. It is widely accepted that the patterns must be *comprehensible*. One of the aspects that are under-addressed in the KDD process is the handling of uncertainty in the process of clustering, classification and association rules extraction. In this paper we present a classification framework for relational databases so as to support uncertainty in terms of natural language queries and assessments. More specifically, we present a classification scheme of non-categorical attributes into lexically defined categories based on fuzzy logic.

## INTRODUCTION

The purpose of Data Mining is the extraction of knowledge from large data repositories. The knowledge may have various forms such as, classifications, association rules, decision trees etc.

In the vast majority of KDD systems and approaches the data values are classified to one of a set of categories that have resulted from a clustering process. Now here we have two issues that may result in knowledge to be partially extracted or not to be extracted at all during the KDD process. We address the following facts and their implications:

- *the clusters are not overlapping*. This means that each database value may be classified into at most one cluster, in some cases it falls out of the cluster limits so it is not classified at all. Though, everyday life experience leads us to the fact that actually a value may be classified to more than one categories. For instance a male person with height 182cm in Central Europe is considered as of "medium" height as well as "tall" to some degree.
- *the data values are treated equally in the classification process*. In traditional data mining systems database values are classified in the available categories in a crisp manner, i.e. a value either belongs to a category or not. The person of the above example is considered as tall and also another person with height 199cm is also considered tall. It is profound that the second person satisfies to a higher degree, than the first, the criterion "tall". This piece of knowledge (the difference of belief that A is tall and also B is tall) cannot be acquired using the schemes.

As it is clear from above brief analysis there is interesting knowledge that is not captured due to the fact that uncertainty is not considered in KDD process.

The KDD process mainly aims at searching for interesting instances of patterns in data sets. It is widely

accepted that the patterns must be *comprehensible* i.e. they should be understood by the analysts[2][6]. This will be achieved by classifying the data into understandable categories represented by natural language values. These natural language expressions should be mapped to the underlying database through a layer that maps the natural language terms to the underlying database schema and values.

Another requirement addressed is the usage and reveal of uncertainty in this context is an important issue [3]. Another issue addressed in the bibliography is the relatively few efforts that have been devoted to classical data analysis techniques like clustering & classification in the area of data mining research [1]. In this paper, we present a classification framework for large relational databases so as to support uncertainty in terms of belief measures. More specifically, we present a scheme that classifies non categorical attribute values into categories maintaining the classification belief. We use fuzzy logic in order to represent and manipulate this belief. We exploit these beliefs in order to support decision making related to one or multiple data sets.

## CLASSIFICATION SCHEME

The term classification implies the procedure according to which each of a set of values is decided to belong into one of a set of related categories. As it is well known, in order to classify a data set there has to be a set of clusters as a result of a preceding clustering process. As we mentioned in previous sections each value that belongs to a cluster (category) should not be treated equally but to contribute according to its classification belief. Thus, we also assume a set of mapping functions assigned to the clusters as a result of an enhanced clustering process described in [5].

### Classification Space (CS)

The term Classification Space (CS) implies the specifications for mapping database values to the fuzzy domain. Each value of the database is classified in one of the above mentioned clusters with an attached d.o.b.

There is inherent uncertainty in the classification of a value in a set of categories. A fundamental issue is the acquisition of the related knowledge i.e.: the categories, the corresponding value ranges, and the mapping functions between the real values and the fuzzy domain. Assuming the appropriate set of value domains for these categories, for each attribute  $A_i$  we define the corresponding *classification set*  $L_{A_i} = \{ct \mid ct \text{ is a classification tuple}\}$ . The classification tuples are of the form:  $(l_i, [v_1, v_2], f_i)$  where  $l_i$  is a lexical category,  $[v_1, v_2]$

is the corresponding value interval and  $f_i$  the assigned transformation function. The value domains may be overlapping. This gives expressive power to the classification mechanism since some values may be classified to more than one categories with different d.o.b.s. Then the collection of all the classification tuples  $ct$  related to the relational schema  $R$  form the Classification Space (CS), which defines the mapping of the data set to the fuzzy domain.

For each tuple  $t_k$  in the data set  $S$  there is a value  $t_{k,A_i}$  that corresponds to the attribute  $A_i$ . Then the d.o.b. that this value belongs to the sets, denoted by the categories and the corresponding domains, is:

$$\mu_{li}(S.t_k.A_i) = f(t_k.A_i) \quad (1)$$

where  $f$  is the transformation function that maps the value  $t_k.A_i$  to the fuzzy domain. Thus, it is clear that for each value  $t_k.A_i$  a set of d.o.b.s  $\{\mu_{li}(S.t_k.A_i)\}$  is produced. Assume  $n_r$  is the number of tuples in the relation and  $l_{A_i}$  is the number of categories corresponding to the attribute  $A_i$ , then the overall number of do.b.s produced is:

$$\sum_{A_i} n_r * l_{A_i} \quad (2)$$

### Classification Value Space (CVS)

The result of the transformation of the data set values to the fuzzy domains using the CS is a 3D structure (see **Figure 1**), further called Classification Value Space (CVS). The front face of this structure stores the original data set while each of the other cells  $C[A_i, l_j, t_k]$ , where  $j, k > 1$ , stores the d.o.b.  $\mu_{li}(S.t_k.A_i)$ .

The higher the d.o.b. is the higher is our confidence that the specific value belongs to the specific set. It is interesting to have an overall measure of classification information, which is included in the values of the attribute with regards to each category.

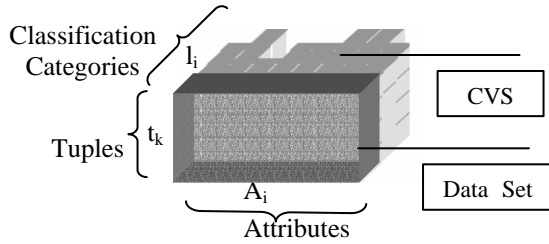


Figure 1. The CVS holding the “degrees of belief” (d.o.b.s) for the classification of the attributes’ values

### Information measures in the CVS

The CVS conveys significant knowledge included in cumulative information measures. One of the important information measures that have been proposed in the bibliography is the Energy metric [4], which reflects the information quantity that is included in a fuzzy set. This information quantity essentially is a measure of the overall belief for a fuzzy set. In sequel, we define various classification quality measures based on energy metric function:

- **Category energy metric**

Let  $A_i$  be an attribute and  $l_i$  a related category. Then the overall belief that the current data set  $S$  contains data that are successfully classified in the category  $l_i$  is given by the normalized information measure:

$$E_{li}(S.A_i) = \left( \sum_k [\mu_{li}(S.t_k.A_i)]^q / n_r \right)^{1/q} \quad (3)$$

where  $n_r$  is the number of tuples in the data set, hence the number of values of the attribute and  $q$  a positive integer. The usual value of  $q$  is 2. Higher values suppress lower d.o.b. making thus the contribution of the tuples with high (close to 1) d.o.b.s more significant. The exponent  $1/q$  is used to amortize the effect of the exponent  $q$ .

We can further compare the information measures of different categories of the same attribute. So for a given data set  $S$  and a given attribute  $A$  with attached categories  $l_1, l_2, \dots, l_n$  the corresponding information measures  $E_{li}(S.A)$  are ordered making thus feasible to decide which category has better support by the data set and also to compare.

- **Attribute energy metric**

The *overall energy* of an attribute  $A_i$ , is the normalized sum of the energy metric values for all the attribute categories. This measure expresses the average overall information energy that is included in the values of the attribute (i.e. how strong is the belief for the classification assessment) and also the amount of information regarding the considered categories. Hence:

$$E_{A_i}(S) = \sum_{l_i} E_{li} / C \quad (4)$$

where  $C$  is the number of categories for the attribute  $A_i$ . Essentially  $E_{A_i}(S)$  is a measure of how successful is the classifications scheme.

- **CVS Energy & classification quality**

The overall information that is included in the CVS and represents the amount of classification information included in the data set is given by the formula:

$$E_{CVS} = \sum_{A_i} E_{A_i} \quad (5)$$

where  $A_i$  are the attributes. This measure is used to compare different data sets as for their information content. Data sets with higher  $E_{CVS}$  contain higher overall measure of information. This energy shows how significant is the information contained in the values of this attribute and also how well the data set fits to the classifications scheme. Also this measure is an indication of the quality of classification. In principle an ideal classification scheme should maximize the  $E_{CVS}$  value.

$E_{CVS}$  is used to compare different datasets that obey to the same schema. An alternative is to use weights in the case that some attributes are more important than others for our purposes. In this case the  $E_{CVS}$  is given by the formula:

$$E_{CVS} = \sum_{A_i} w_i E_{A_i} \quad (5a)$$

where  $0 \leq w_i \leq 1$  is the weight attached to the attribute  $A_i$ .

### Composite classifications

Another need that arises is the representation of the d.o.b. related to composite classifications of tuples. We adopt the min measure for composition of fuzzy predicated from the bibliography [Fuzzy book]. Thus for two attributes  $A_t$ ,  $A_e$  and  $l_i$ ,  $l_j$  two corresponding categories the d.o.b. that a tuple  $t_k$  belongs to the set, characterized by the predicate “ $A_t.l_i$  and  $A_e.l_j$ ”, is given by the equation:

$$\mu_{l_i \text{ and } l_j}(t_k.A_t, t_k.A_e) = \min(\mu_{l_i}(t_k.A_t), \mu_{l_j}(t_k.A_e)) \quad (6)$$

In an analogous to the previous the overall information measure related to the criterion “ $A_t.l_i$  and  $A_e.l_j$ ” is given by the equation:

$$E_{l_i \text{ and } l_j}(A_t, A_e) = \left( \sum_k \left[ \mu_{l_i \text{ and } l_j}(t_k.A_t, t_k.A_e) \right]^q / N \right)^{1/q} \quad (7)$$

which represents the belief that tuple  $t_k$  has both features (belongs to both categories)  $l_i$ ,  $l_j$  and, therefore, it is classified accordingly.

### Reasoning with information measures

The result of the KDD procedure is a set of assessments about the underlying data. These assessments should be in an understandable form for the humans so that they will be useful and exploitable. The scheme presented above contributes to this requirement. The information measures mentioned above are exploited to support queries and decision support of the following categories:

- *Single data set, single attribute queries*

Here we have queries related to categories of an attribute in the same data set. In Table 1 there is a list of indicative queries and the way they are handled by the classifications scheme.

Query	Value returned
“Which is the most important group of salaries in the data set?”	Max( $E_{l_i}(S.salary)$ )
“Does the attribute salary include mostly high or medium salaries?”	if( $E_{high}(S.salary) > E_{medium}(S.salary)$ ) return $E_{high}(S.salary)$ else return $E_{medium}(S.salary)$

Table 1. Queries and the related information measures.

- *Multi-dataset queries*

In this category we are concerned with queries that involve more than one data sets of the same relational schema and CS. Assume two data sets including sales in two different supermarkets namely S1, S2, then the following query can be processed using the information measures as appears in Table 2.

### CONCLUSIONS

One of the objectives of a KDD process is to produce understandable knowledge in terms of patterns detected in a large data set.

In this work we presented a scheme for classification of

Query	Value returned
“Which of the S1, S2 contains more transactions made early morning?”	if ( $E_{morning}(S1.time\_of\_p) > E_{morning}(S2.time\_of\_p)$ ) return $E_{morning}(S1.time\_of\_p)$ else return $E_{morning}(S2.time\_of\_p)$

Table 2: Queries involving multiple data sets

database values into categories characterized by lexical values. The classification maintains the uncertainty of the classification through the maintenance of a framework based on fuzzy logic. This scheme enables moreover the evaluation of accumulated information measures based on the energy metric function. Thus we can compare different sets as to the degree they fit to the classification scheme. More specifically the energy metric for a category can be used to compare different datasets under the specific criterion.

Further work will be concentrated in usage of the proposed framework in order to adjust an original classification model according to the feedback we get by classifying different datasets. This adjustment will result in classification and association schemes that maximize the energy metrics functions related to the various related entities. The overall objective in this case is the incremental production of optimal natural language classification and association extraction models.

The Ecvs information measure can be used to evaluate a classification scheme over a set of data sets. For instance when  $E_{A_i}$  is maximized we have strong evidence that the classification scheme fits well the data set. This repeated over many data sets (of the same schema) and by adjustment of the parameters of the classification scheme (number and kind of categories, domain values, and transformation functions) results in an optimum classification system for the specific data sets.

### REFERENCES

- [1] Chaudhuri S. (1997), Data Mining and Database Systems: where is the intersection, *bulleting of the IEEE CS TC on Data Engineering*.
- [2] Fayyad U. Piatetsky-Shapiro G, Smyth P(1996), The KDD process for extracting Useful Knowledge from Volumes of Data, *CACM* 39(11), 27-35
- [3] Glymour C., Madigan D., Pregibon D, Smyth P. (1996), Statistical Inference and Data Mining, *CACM* 39(11), 35-42
- [4] Gupta, M., and Yamakawa, T., (1988), (eds), *Fuzzy Logic and Knowledge Based Systems, Decision and Control* (North-Holland).
- [5] M. Halkidi, M. Vazirgiannis (1999), Clustering: Quality measures and uncertainty handling. *Technical report*, Athens Univ. of Economic & Business.
- [6] Kloegen W. (1996), Explora: AI Multipattern and Multistrategy Discovery Assistant, in: U. Fayyad, et al.(eds), *Advances in Knowledge Discovery and Data Mining*, AAAI Press.