

Uncertainty handling in the data mining process with fuzzy logic

Michalis Vazirgiannis, Maria Halkidi

Department of Informatics

Athens University of Economics & Business

Patision 76, 10434, 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 and provides decision support facilities based on related information measures.

I. 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. Then, 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 into more than one categories. For instance a male person 182cm high 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 199cm high 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 the above brief analysis there is interesting knowledge that is not captured due to the fact that uncertainty is not considered in the KDD process.

The KDD process mainly aims at searching for interesting instances of patterns in data sets. It is widely accepted that the patterns (e.g. classifications, rules etc) must be *comprehensible* i.e. they should be understood by the analysts [8][5]. Assume the transaction log of a computer sales store, and that a subset of its scheme is: $R = \{client_salary, client_age, price\}$. Applying the techniques proposed in [12], we would have to come up with rules of the form:

$client_salary[8000,11000] \text{ and } client_age[25-40] \Rightarrow price[1300,2000]$

Apparently the rule introduced above is not clearly comprehensible, since it does not place the rule in the greater context of the involved attributes (i.e. what does that range $client_salary[3000,4500]$ mean in the full range of salaries as well as in their population distribution features?). A manager/analyst, as non-domain expert, would not understand the meaning of such a rule since the underlying data semantics are not made clear in the rule context. Thus, a requirement for understandable patterns of knowledge as results of the data mining process arises. This will be achieved by classifying the data into understandable categories represented by natural language values.

Another issue is the “crispness” of the value domains imposed by this approach. For instance, (see Table I), the tuple with $tid=11$ is excluded from the supporting set although all its values support quite well the rule apart from the value of the attribute “price” which is only 0.00615% out of the required range. The result is that many “interesting” tuples (i.e. contributing to the semantics hidden behind such a rule), are rejected due to the crisp limits that have been set. It is evident that the problem here is that the classification of the values in these domains is flat, i.e. all the values in the domain are treated equally as for the criterion of partitioning (i.e. the price). The partition in domains reflects the classification of the attribute values in categories (i.e. “very cheap”, “cheap”, “moderate”, “expensive”). 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.

TABLE I

THE SALES TRANSACTION LOG TABLE

Tid	Client_salary	client_age	Price
1	6387	64	567
2	4048	70	261
3	5829	53	307
4	6576	60	166
5	7832	46	1169
6	8243	54	713
7	9218	21	1458
8	3857	76	1038
9	5030	22	681
10	4447	19	136
11	9765	36	1292
12	6822	37	1136
13	8763	79	1444
14	1643	66	8
15	5387	73	283
16	2943	71	173
17	4584	76	641
18	6963	69	983
19	2323	80	742

Another requirement addressed is the usage and reveal of uncertainty in this context is an important issue [6]. 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 [2].

In this paper we propose a methodology that represents uncertainty in the classification stages in KDD environment 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.
- Information measures for the evaluation of the above defined classification scheme based on fuzzy logic concepts. We can exploit classification belief based on these measures in order to support decision-making related to one or multiple data sets.

The paper is organized as follows. In section two we present the classification scheme while in section three we propose information measures based on fuzzy logic in order to evaluate and exploit the information included in proposed scheme. In section four we elaborate on the multi-dimensional extension of this scheme, while in section five we present reasoning based on proposed information measures. We conclude in section six by summarizing and providing further work directions.

II. 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. In this research effort we assume that there is a given set of clusters for each attribute. As we mentioned in previous sections each value that belongs to a cluster (category) should not be treated equally but 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. Then each database value is mapped to a category bearing a d.o.b. (degree of belief) for this classification as a result of using the corresponding mapping function.

The classification scheme is applied on a data set S under a certain relational schema $R = \{A_i\}$ where A_i is an attribute. The values of the non-categorical attributes (A_i) are classified into categories according to a set of categories $L = \{l_i\}$ (where l_i a category, for instance: "tall", "short" etc.) and a set of classification functions based on fuzzy logic methodologies. The result of this procedure is a set of degrees of belief (d.o.b.s) $M = \{\mu_{li}(t_k.A_i)\}$. Each member of this set represents the confidence that the specific value $t_k.A_i$ (where t_k is the tuple identifier) belongs to the set denoted by the category l_i .

A. Classification space (CS)

The term Classification Space (CS) implies the specifications for mapping data base values to the fuzzy domain. Each value of the database is classified in one of the above mentioned categories (clusters) with an attached d.o.b. We assume the attribute "client_salary" from our running example, which in a data set ranges between the values 1500 and 10000. In real world people characterize the value of a salary as *high*, *low*, *moderate*. How would one classify a specific salary value

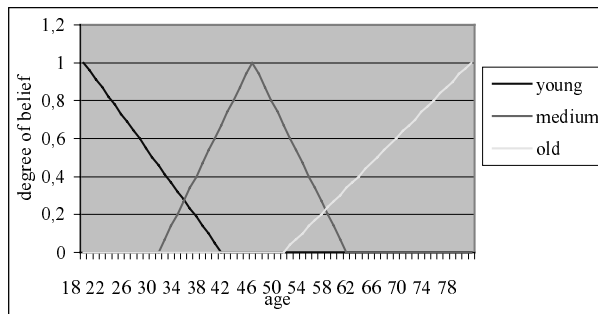


Fig.1.The transformation functions for the attribute "client_age"

into a category? What are the values' ranges corresponding to these categories? Are they overlapping?

As it is clear, 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 increases the expressive power of the classification mechanism since some values may be classified to more categories than one with different d.o.b.s. Then the collection of all the classification tuples ct related to the relational schema R forms the Classification Space (CS), which defines the mapping of the data set to the fuzzy domain.

In Table II the CS appears based on the schema of our sales example (see Table I). For each attribute, a set of lexical categories, the corresponding domain limits and the related transformation functions are provided.

The transformation function selection is an important issue that can affect the results of classification. In our system we have currently adopted linear functions (decreasing, triangle and increasing) [7]. In Figure 1 the mapping of the age values to the fuzzy domain appears based on the CS specifications (see Table II).

TABLE II.
THE CLASSIFICATION SPACE FOR THE SALES SCHEMA.

client_salary				
	low	Medium	High	
Min	1500	2500	4000	
Max	3000	5500	10000	
Function	decr	triangle	increasing	

client_age			
	young	Medium	Old
Min	18	30	50
Max	40	60	80
Function	decr	triangle	increasing

price				
	very cheap	Cheap	moderate	expensive
Min	1	10	35	70
Max	15	50	80	150
Function	decr	triangle	triangle	triangle

As it is depicted in the figure 1 the value domains may be overlapping, so that an age value may be classified into two categories. In a similar way the rest of the attributes are mapped 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. The choice of functions is a fundamental issue and will have a great impact on the creditability of the d.o.b.s. 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 d.o.b.s produced is:

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

B. 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 2) further called Classification Value Space (CVS). The front face of this structure stores the original data set (included in Table I) while each of the other cells $C[A_i, l_j, t_k]$, where $j, k > 1$, stores the d.o.b. $\mu_{li}(S.t_k.A_i)$. Further we reference a cell in the CVS as $CVS(t_k.A_i.l_j)$.

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 regard to each category.

The algorithm for computing the d.o.b.s for the data set values with reference to the CS follows:

```

for each attribute  $A_i$  in CS
  for each category  $C_j$  of  $A_i$ 
    for each value  $t_k.A_i$  in the data set
      compute d.o.b.( $A_i, C_j, t_k.A_i$ )
    end
  end
end

```

Thus the time complexity is $O(d*c*n)$ where d is the number of attributes (data set dimension), c is the number of categories (clusters) and n is the number of d.o.b. values for a category n (number of tuples in the data set). Usually $c, d \ll n$. Thus, the time complexity for computing the d.o.b.s for a data set will be $O(n)$.

III. 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 [7], 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.

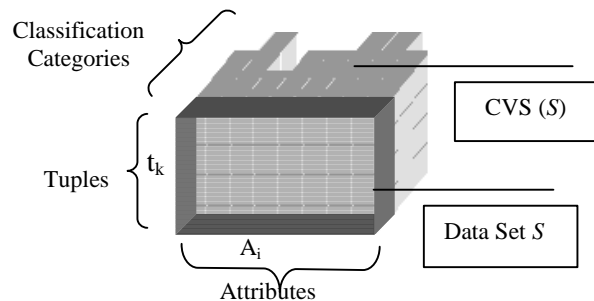


Fig. 2. The CVS holding the “degrees of belief” (d.o.b.s) for the classification of the attributes’ values.

A. 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:

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

where n_r is the number of tuples in the data set, hence the number of values of the attribute and q is 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. For instance the query: “Does the sales database contain mostly high of medium salaries?” is answered by comparing the values: $E_{medium}(salary), E_{high}(salary)$ as resulting from (3).

B. 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:

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

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.

C. 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 equation:

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

where A_i are the attributes. The result is the information content of the CVS. This measure is used to compare different data sets as for their information content. Data sets with higher E_{CVS} correspond to 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 classification scheme. Also this measure is an indication of the quality of classification. In principle an ideal classification scheme should maximize the E_{cvs} value. Indeed when E_{cvs} is maximized the uncertainty is minimized and thus the confidence for classification is high.

IV. MULTI-DIMENSIONAL CLASSIFICATIONS

Another need that arises is the representation of the d.o.b. related to composite classifications of tuples. For instance we are interested to know to what degree a tuple in our sample data set satisfies more than one criteria e.g.: “*morning and cheap purchases*”. The term “*morning and cheap*” defines a new category and we need to provide a mapping function for this. In this case we can introduce two alternatives:

- *Classification based on multi-dimensional clusters.* In this case we define clusters (initial categories) for our data set taking into account all the attributes referred to our criteria. Then the clustering process produces multi-dimensional clusters and we can define the membership functions for them based on the procedure used in the case of one-dimensional data sets.
- *Classification based on one-dimensional clusters.* We adopt the min measure for composition of fuzzy predicates from the bibliography [7]. Thus for two attributes A_t , A_e and l_i , l_j two corresponding categories (l_i referring to A_t and l_j to A_e), 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:

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

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

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

which represents the belief that tuple t_k has both features (belongs to both categories) li , lj and, therefore, it is classified accordingly. For instance, we may submit the query: “What is the overall belief that the database contains transactions for cheap purchases made in the morning?”.

A. An experimental study of multi-dimensional classification approaches

The objective of this study is to compare two approaches described in section III for the definition of multi-dimensional classification. More specifically, we use a data set with data related to stock exchange transactions and we compute the overall energy produced by the adoption of the above-described alternatives. The size of our data set was 1000 tuples and its schema is : $R = \{close_price, high_price, volume\}$, where *close_price* is the daily closing price of the stock, *high_price* is the highest price of the stock during a session and *volume* is the number of transactions for the specific stock. In this point, we also have to mention that the energies of our data set classification scheme computed based on a system we have implemented according to the above described classification framework.

TABLE III
ENERGY METRIC FOR A GIVEN NUMBER OF CLUSTERS

Close Price, Volume

<i>One-dimensional clustering</i>							
	<i>C1</i>	<i>C2</i>	<i>C3</i>	<i>C4</i>	<i>C5</i>	<i>C6</i>	<i>C7</i>
<i>Category</i>	0,176	0,834	0,068	0,122	0,361	0,022	0,049
<i>Energy</i>	<i>C8</i>	<i>C9</i>					
	0,221	0					
<i>Ecl_vol</i>	0,206						

<i>Two-dimensional clustering</i>							
	<i>Cat1</i>	<i>Cat2</i>	<i>Cat3</i>	<i>Cat4</i>	<i>Cat5</i>	<i>Cat6</i>	<i>Cat7</i>
<i>Category</i>	0,174	0,176	0,055	0,138	0,379	0,352	0,319
<i>Energy</i>	<i>Cat8</i>	<i>Cat9</i>					
	0,669	0,141					
<i>Ecl_vol</i>	0,266						

TABLE IV
ENERGY METRIC FOR THE OPTIMAL CLUSTERING SCHEME.
Close Price, Volume

<i>One-dimensional clustering</i>							
	<i>C1</i>	<i>C2</i>	<i>C3</i>	<i>C4</i>	<i>C5</i>	<i>C6</i>	<i>C7</i>
<i>Category</i>	0,117	0,164	0,011	0,086	0,569	0,299	0,222
	<i>C8</i>	<i>C9</i>	<i>C10</i>	<i>C11</i>	<i>C12</i>	<i>C13</i>	<i>C14</i>
	0,101	0,078	0,001	0,033	0,194	0,146	0,133
	<i>C15</i>	<i>C16</i>	<i>C17</i>	<i>C18</i>	<i>C19</i>	<i>C20</i>	<i>C21</i>
<i>Energy</i>	0,012	0,124	0	0,002	0,141	0,069	0,101
	<i>C22</i>	<i>C23</i>	<i>C24</i>	<i>C25</i>	<i>C26</i>	<i>C27</i>	<i>C28</i>
	0,064	0,122	0,045	0,065	0,31	0,221	0,104
	<i>C29</i>	<i>C30</i>	<i>C31</i>	<i>C32</i>	<i>C33</i>	<i>C34</i>	<i>C35</i>
<i>Ecl_vol</i>	0,126						

<i>Two-dimensional clustering</i>							
	<i>Cat1</i>	<i>Cat2</i>	<i>Cat3</i>	<i>Cat4</i>	<i>Cat5</i>	<i>Cat6</i>	<i>Cat7</i>
<i>Category</i>	0,174	0,176	0,055	0,138	0,391	0,352	0,319
<i>Energy</i>	<i>Cat8</i>						
	0,669						
<i>Ecl_vol</i>	0,284						

The overall result of this study is that the first approach based on multi-dimensional clusters, produces better classification schemes. Multi-dimensional clustering extracts clusters that are the best partitioning for a data set as it examines simultaneously all the attributes (dimensions). Also categories that are not supported by the data set ignored and thus the classification scheme could be adjusted better to the data set. Assuming a two dimensional data set (*close_price, volume*) of stock exchange database, we demonstrate the above with the following experiments:

1. *Classification scheme based on a given number of clusters.* In this case, the clustering procedure is applied for a given number of clusters so as to compare the results of the two approaches with respect to the definition of a data set partitioning that is as good as possible for the given number. More specifically, we apply clustering to each of the attributes (*close_price, volume*) so as to define three clusters (categories) for each of them. Thus, we defined nine new clusters for the category “*close_price* and “*volume*” combining the extracted categories of each attribute. Then we apply two-dimensional clustering in order to define a partitioning of the data set into nine clusters. Table III presents the overall energy (as in (4)) as computed in each of the approaches. As it is obvious, the overall belief produced by two-dimensional clustering is higher. It is also noteworthy that, none of the nine categories produced by multi-dimensional clustering has zero (0) energy in contrast to the case of one-dimensional. As a consequence, the approach based on multi-

dimensional clustering searches for the best nine clusters that can be extracted by the data set and ignores the categories that are not supported.

2. *Classification based on optimal clustering schemes.* In this case, we apply clustering procedure giving a range in which the number of clusters can take values and we ask for the optimum clustering scheme. The selection of the clustering scheme is based on well-defined quality clustering criteria. Table IV shows the overall energy in each of the multi-dimensional classification approaches. The result of comparing two approaches is that the overall belief produced by two - dimensional clustering is higher and as a consequence the classification scheme defined is better. The clustering procedure has defined the optimum partitioning of the data set taking into account both attributes and thus the outcome (clusters) are adjusted better to our data set than the clusters produced by the combination of clusters defined by separate attributes.

We carried out a similar study for three-dimensional data sets and we concluded into similar results to two-dimensional data i.e. that multi-dimensional clustering could result in better initial categories for the multi-dimensional classification

V. 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, since the results of the data set can be represented in the form of natural language statements. The information measures mentioned above are exploited to support queries and decision support of the following categories:

A. Single data set, single attribute queries

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

B. Multi-data set queries

In this category we are concerned with queries that involve two or more data sets of the same relational schema and CS. Assume two data sets including sales in two different supermarkets namely S1, S2. Then the queries appeared in Table VI can be processed using the information measures defined above.

VI. RELATED WORK

One of the three components of a KDD system is the *model* whose functions among others include [5] the classification procedure. "Classification" aims at mapping an object to a predefined set of categories/classes, unlikely to the "clustering" procedure where the extraction of the classes from a set of data is achieved by finding grouping of values and similarity metrics.

The classification problem has been studied extensively in statistics, pattern recognition and machine learning community as a possible solution to the knowledge acquisition or knowledge extraction problem [11].

TABLE V.
SAMPLE QUERIES AND THE RELATED INFORMATION MEASURES.

Query	Value returned
"What is the belief that the data set contains high salaries?"	$E_{\text{high}}(\text{S.salary})$
"Does the attribute salary include mostly high or medium salaries?"	if ($E_{\text{high}}(\text{S.salary}) > E_{\text{medium}}(\text{S.salary})$) return $E_{\text{high}}(\text{S.salary})$ else return $E_{\text{medium}}(\text{S.salary})$

TABLE VI
QUERIES INVOLVING MULTIPLE DATA SETS

Query	Value returned
"Which of the S1, S2 contains more transactions made early morning?"	if ($E_{\text{morning}}(\text{S1.time_of_p}) > E_{\text{morning}}(\text{S2.time_of_p})$) return $E_{\text{morning}}(\text{S1.time_of_p})$ else return $E_{\text{morning}}(\text{S2.time_of_p})$
"In which supermarket there are more cheap purchases made in the evening?"	if ($E_{\text{cheap and evening}}(\text{S1.price, S.time_of_p}) > E_{\text{cheap and evening}}(\text{S2.price, S.time_of_p})$) return $E_{\text{cheap and evening}}(\text{S1.price, S.time_of_p})$ else return $E_{\text{cheap and evening}}(\text{S2.price, S.time_of_p})$

A number of classification techniques have been developed and are available in bibliography. Among these, the most popular are: *Bayesian classification* [3], *Neural Networks* [1] and *Decision Trees* [14].

The above reference to some of the most widely known classical classification methods denotes the relatively few efforts that have been devoted to data analysis techniques (i.e. classification) in order to handle uncertainty. These approaches produce a crisp classification decision, so an object either belongs to a class or not, which means that all objects are considered to belong to a class equally. Moreover, most of the classification proposals and algorithms consider the classes as non-overlapping [8]. It is obvious that there is no notion of uncertainty representation in the proposed methods, though usage and reveal of uncertainty is recognised as an important issue in research area of data mining [6]. For this purpose, the interest of research community has been concentrated on this context and new classification approaches have recently been proposed in bibliography so as to handle uncertainty.

The issue of classification involves the definition of categories that group the values of an attribute *A* in sets that have a specific feature. A recent approach in classification for data mining is presented in [4].

Also, an important issue in data clustering and classification is the extraction of appropriate value intervals that correspond to logical categories related to an attribute. An interesting approach related to this issue is addressed in [9].

An approach for pattern classification based on fuzzy logic is represented in [10]. The main idea is the extraction of fuzzy rules for identifying each class of data. The rule extraction methods are based on estimating clusters in the data and each cluster obtained corresponds to a fuzzy rule that relates a region in the input space to an output class. Thus, for each class c_i the cluster centre is defined that provides the rule: *If {input is near x_i } then class is c_i .* Then for a given input vector x , the system defines the degree of fulfilment of each rule and the consequent of the rule with highest degree of fulfilment is selected to be the output of the fuzzy system. As a consequence, the approach uses fuzzy logic to define the best class in which a data value can be classified but the final result is the classification of each data to one of the classes.

In [13], an approach based on fuzzy decision trees is presented and aims at uncertainty handling. It combines symbolic decision trees with fuzzy logic concepts so as to enhance decision trees with additional flexibility offered by fuzzy representation. More specifically, they propose a procedure to build a fuzzy decision tree based on classical decision tree algorithm (ID3) and adapting norms used in fuzzy logic to represent uncertainty [13]. However, there is no evaluation of proposed inference procedures as regards the quality of new sample classification.

In general, there are some approaches proposed in bibliography, which aim at dealing with uncertainty representation (e.g. fuzzy decision trees). According to these approaches each data value can be assigned to more than one categories with an attached degree of belief. However, they don't propose ways to handle classification information and exploit it for decision-making. In this paper, we propose an approach that aims at uncertainty handling in the classification process based on fuzzy logic concepts. We propose a classification framework that maps data to fuzzy domains and maintains uncertainty in terms of degrees of belief.

VII. CONCLUSIONS

One of the objectives of a KDD process is to produce understandable knowledge in terms of patterns detected in a large data set. We feel there is a lot of potential in the area of mining patterns of knowledge, as regards classification of quantitative attributes. In this paper we presented:

- A scheme for classification of database values putting emphasis in uncertainty handling and classification quality measures. The classification scheme maintains the uncertainty through the maintenance of a framework based on fuzzy logic.
- Information measures for the classification scheme based on the energy metric function, which reflect the information quantity that is included in a fuzzy set. Based on these measures, we can compare different data sets as to the degree they fit to the classification scheme or compare different data sets under a specific criterion. Also, we extract "useful" knowledge for reasoning and decision making based on the information measures.

Moreover, we present how the proposed classification scheme can be used for multi-dimensional classification so as to support decision-making that combines more than one-classification criteria. For this purpose, we proposed two approaches: i) *classification based on multi-dimensional clusters*, ii) *classification based on one-dimensional clusters*, while we described an experimental study we have carried out in order to evaluate these two approaches. The overall result of this study is that the approach based on multi-dimensional clusters, produces better classification schemes.

Further work will be concentrated in usage of the proposed framework in order to adjust an initial classification model according to the feedback we get by classifying different data sets. This adjustment will result in classification scheme that maximizes the energy metric functions related to the various related entities. The overall objective in this case is the incremental production of optimal classification and association extraction models. Also, we aim at the study of different mapping functions and their effect to the proposed classification scheme as regards uncertainty representation. Moreover, in future, more information measures for our

classification scheme will be proposed based on various proposed in bibliography, and they will be evaluated in order to select the optimal definition for the classification quality measures.

REFERENCES

- [1] M. Berry, G. Linoff. *Data Mining Techniques For marketing, Sales and Customer Support*. John Willey & Sons, Inc, 1996.
- [2] S. Chaudhuri, "Data Mining and Database Systems: where is the intersection", bulleting of the IEEE CS TC on Data Engineering, 1997
- [3] P. Cheesman, J. Stutz, "Bayesian Classification (AutoClass): Theory and Results" in *Advances in Knowledge Discovery and Data Mining* (Editors: U. Fayyad, et al.), AAAI Press, 1996.
- [4] M. Dalkilic, E. Robertson, D.V. Gucht, "CE: The classifier-Estimator for Data Mining", in the proceedings of IFIP-DS7 Conference on Database Semantics, 1997
- [5] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, "The KDD process for extracting Useful Knowledge from Volumes of Data", in *CACM* vol.39 (11), 1996, pp. 27-35
- [6] C. Glymour, D. Madigan, D. Pregibon, P. Smyth, "Statistical Inference and Data Mining", in *CACM* vol.39 (11), 1996, pp. 35-42
- [7] M.GUPTA, and T. YAMAKAWA, (eds), *Fuzzy Logic and Knowledge Based Systems, Decision and Control* (North-Holland), 1988.
- [8] W. Kloegen, "Explora: AI Multipattern and Multistrategy Discovery Assistant", in the book "Advances in Knowledge Discovery and Data Mining" (Editors: U. Fayad, et al.), AAAI Press, 1996.
- [9] D. Rasmussen, R. Yager, "Induction of Fuzzy Characteristic Rules", in the proceedings of the First European Symposium PKDD, Trondheim, 1997
- [10] S. Chiu. "Extracting Fuzzy Rules from Data for Function Approximation and Pattern Classification". *Fuzzy Information Engineering- A Guided Tour of Applications*. (Eds.: D. Dubois, H. Prade, R Yager), 1997.
- [11] R. Rastori, K. Shim. "PUBLIC: A Decision Tree Classifier that Integrates Building and Pruning". *Proceeding of the 24th VLDB Conference*, New York, USA, 1998.
- [12] R. Srikant, R. Agrawal, "Mining Quantitative Association Rules in Large Relational Tables", in the proceedings of ACM-SIGMOD '96 Conference.
- [13] Z. Cezary, Janikow, "Fuzzy Decision Trees: Issues and Methods", *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 28, Issue 1, pp 1-14, 1998.
- [14] T. Mitchell. *Machine Learning*. McGraw-Hill, 1997