

Managing uncertainty and quality in the classification process

Maria Halkidi, Michalis Vazirgiannis
Dept of Informatics, Athens University of Economics & Business
76 Patission Str, 10434
Email: {mhalk, mvazirg}@aueb.gr

Abstract. An important open issue in KDD research is the reveal and the handling of uncertainty. The popular classification approaches do not take into account this feature while they do not exploit properly the significant amount of information included in the results of classification process (i.e., classification scheme), though it will be useful in decision-making. In this paper we present a framework that maintains uncertainty throughout the classification process by maintaining the classification belief and moreover enables assignment of an item to multiple classes with a different belief. Decision support tools are provided for decisions related to: i. relative importance of classes in a data set (i.e., “young vs. old customers”), ii. relative importance of classes across data sets iii. the information content of different data sets. Finally we provide a mechanism for evaluating classification schemes and select the scheme that best fits the data under consideration.

1 Introduction and Motivation

Classification is one of the main tasks in the data mining procedure for assigning a data item to a predefined set of classes. According to [4], *classification* can be described as a function that maps (classifies) a data item into *one* of the several predefined classes.

A well-defined set of classes and a training set of pre-classified examples characterize the classification. On the contrary, the clustering process does not rely on predefined classes or examples[1]. The goal in classification process is to induce a model that can be used to classify future data items whose classification is unknown [1, 12]. For this purpose, many classification approaches have been developed and are available in literature. However, in the vast majority of traditional approaches the data values are classified to one of the classes. Also, the issue of evaluation of classification outcome (i.e., classification scheme) is under-addressed in most of classification approaches. Hereafter we address issues that arise from classification approaches:

- i. *The clusters are not overlapping.* The limits of clusters are crisp and each database value may be classified into at most one cluster. This is unlikely to everyday life experience where a value may be classified into more than one classes (clusters)*. For instance a male person 182cm high in Central Europe is considered as of “medium” height as well as “tall” to some degree.
- ii. *The data values are treated equally in the classification process.* In traditional data mining systems database values are classified in the available classes in a crisp manner i.e., a value either belongs to a category or not. Also, they assume that all values belong to a class, assigned to it with the same degree of belief. The person 182cm high is considered 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 of being “tall”. This piece of knowledge (the difference of belief that A is tall and also B is tall) cannot be acquired using the well-established classification schemes.
- iii. *Classification results may hide “useful” knowledge for our data set.* Most of the classification methods define a model that is used to classify new instances to predefined classes. This assignment of data values to classes conveys significant knowledge and when aggregated for many values provides collective indication on a

* In this paper we use the terms “classes” and “clusters” interchangeably.

class importance. We can exploit this aggregated knowledge for decision-making as well as for selection of the classification model that best fits a data set.

Motivation. Our effort is not yet another classification algorithm for learning (discovering) classes. We address a somewhat different issue. Given a data set S consisting of a set of tuples $\{t\}$ each of which consists of a set of values $\{t.v_i\}$ where v_i corresponds to the value of an attribute A_i , we want to be able to:

- decide to which class(es) a value set v_i can be assigned and what the respective beliefs for each assignment is,
- compare the i. relative importance of classes in a data set (e.g. “young vs. old customers”), ii. relative importance of classes across data sets, iii. the information content of different data sets,
- assess the quality of a classification model i.e., how well it fits a data set. This requirement arises as data sets continuously change due to insertions/deletions/updates.

In recent literature some efforts based on probabilistic concepts [17] deal with uncertainty. They measure the probability for a data set value to belong to a category, while we measure the degree of belief with which it is classified in a category. Our approach is a fuzzy alternative for the classification process that supports the uncertainty using degrees of belief and not probabilities.

In the decision tree family of algorithms successive splits of a data set S into non-overlapping sets $\{S_i\}$ take place. The split is based on the different values of an attribute A_i (selected so that a metric is minimized). Then each successive step splits the subsets in other subsets so that each tuple of the data set is assigned to one class. In each of the splits a crisp decision is made and according to the above discussion potential knowledge is lost. A related approach dealing with uncertainty is fuzzy decision trees [15]. According to it, a data value can be classified to several tree nodes with an attached degree of satisfaction. More specifically, to classify a new value, we should find leaves (i.e., classes) whose restrictions are satisfied by this value. Then we combine the restrictions along the path from root to the specific leaves and their satisfaction degrees into a single crisp response. It is obvious that the classification result of fuzzy decision trees is crisp though they use fuzzy concepts during classification process. Moreover, the classification of a new data value is based on successive tests to internal nodes i.e., it is classified according to one attribute each time with an attached degree of satisfaction. However, as we proved in [19] classification based on multi-dimensional classes (i.e., classes defined taking into account many attributes simultaneously) results in better classification inferences. Even in the fuzzy decision trees approach the eventual assignment is crisp since each item is assigned to the most probable class where as in our approach an item may be classified into different classes. Moreover, the classification belief for each assignment is maintained.

Our contribution. The contributions of this paper are summarized as follows:

- *Maintenance of classification belief* all the way through the classification process and moreover a value set can be assigned to more than one classes with a different belief.
- *Decision support tools* for decision related to: i. relative importance of classes in a data set (e.g., “young vs. old customers”) and ii. relative importance of classes across data sets iii. the information content of different data sets.
- *Quality assessment* of a classification model. This procedure will be very useful for evaluating models and select the one that best fits the data under consideration.

It is important to stress that our contribution is independent of any classification algorithm. Indeed, we take as input the classes resulting from the application of any algorithm on a training set and we classify all the data set to these classes introducing uncertainty features. Moreover we take into account aggregate beliefs that will assist for decision support in the data set and across data sets.

The remainder of the paper is organized as follows. Section 2 surveys related work. Section 3 elaborates on the proposed classification approach while in Section 4 we present the fundamental concepts of the proposed classification framework. In Section 5 we define classification information measures so as to exploit the classification belief. Section 6 presents a quality assessment procedure for a classification scheme while in Section 7 we discuss the results of an experimental study we carried out using synthetic and real datasets. We conclude in Section 8 by summarizing and providing further research directions.

2 Related work

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

Bayesian classification is based on bayesian statistical classification theory. The aim is to classify a sample x to one of the given classes c_1, c_2, \dots, c_N using a probability model defined according to Bayes theory[3]. Also, complete knowledge of probability laws is necessary in order to perform the classification [7].

Decision trees are one of the widely used techniques for classification and prediction. A number of popular classifiers construct decision trees to generate classification models. A decision tree is constructed based on a training set of pre-classified data. Each internal node of the decision tree specifies a test of an attribute of the instance and each branch descending of that node corresponds to one of the possible values for this attribute. Also, each leaf corresponds to one of the defined classes. Some of the most widely known algorithms that are available in literature for constructing decision trees are: ID3[10], C4.5[11], SPRINT[13], SLIQ[9], CART etc.

Another classification approach used in many data mining applications for prediction and classification is based on neural networks [1].

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, that is an object either belongs to a class or not and all objects are considered to belong in a class equally. It is obvious that there is no notion of uncertainty representation in the proposed methods, though usage and reveal of uncertainty is recognized as an important issue in research area of data mining[14]. For this purpose, the interest of research community has been concentrated on this context and new classification approaches have recently been proposed in literature so as to handle uncertainty. In this point we should mention that the issue of uncertainty handling is not restricted to the classification but there is also an important set of clustering approaches that aims at uncertainty handling [17, 18]. However, in this paper we concentrated on the case of classification process.

An approach for pattern classification based on fuzzy logic is represented in [2]. 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 center 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 fulfillment of each rule and the consequent of the rule with highest degree of fulfillment 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.

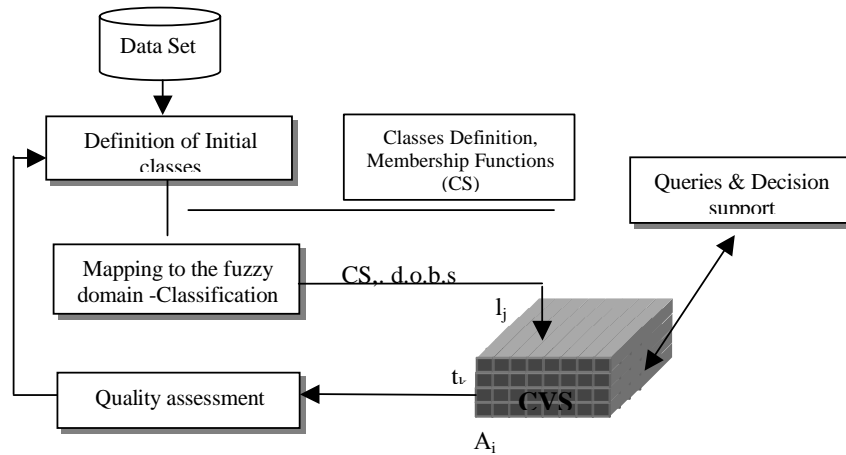


Figure 1. Steps of the classification approach.

In [15], 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 [15]. As a consequence, the tree-building procedure is the same as that of ID3. The difference is that a training example can be partially classified to several tree nodes. Thus, each instance of data can belong to one or more nodes with different membership that is calculated based on the restriction along the path from root to the specific node. However, according to the decision tree methodology the classification inferences are crisp. More specifically, to define the classification assigned to a sample, we should find leaves whose restrictions are satisfied by sample and combine their decisions into a single crisp response. Furthermore, there is no evaluation of proposed inference procedures as regards the quality of new sample classification. Also, there is a significant amount of information included in decision tree that is not exploited and thus there is useful knowledge that is not extracted.

In general, there are some approaches proposed in literature, which aim at dealing with uncertainty representation (e.g. fuzzy decision trees). According to these approaches each data value can be assigned to one or more classes with an attached degree of belief. However, they don't propose ways to handle classification information and exploit it for evaluation of classification scheme and decision-making. Another related issue is how well a classification model fits an evolving data set. As new data values are inserted to the data set, it is possible the statistical features of the classes to be affected and then the classification model should be updated. It is obvious that there is a need to define an evaluation procedure for classification schemes, which helps us to understand how successful the classification for a specific data set is. However, the evaluation of classification models is under-addressed by most classification approaches where as we address and tackle the issue.

3 System Architecture

As it is well known, the classification procedure is based on a predefined set of classes. We assume a set of classes as a result of a preceding clustering process, which aims at the definition of the “optimal” clustering scheme that fits a specific data set [6]. More specifically, we apply a clustering algorithm that results in a clustering scheme

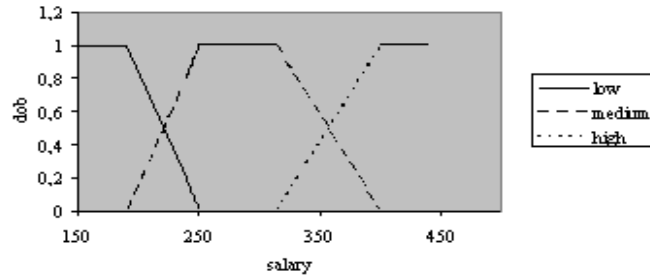


Figure 2. The transformation function (HMF) for one-dimensional data set.

corresponding to the initial classes on which the classification process is based. However, the majority of clustering algorithms result in crisp clusters (i.e., they assume that each data item belongs to only one cluster), but as we have mentioned in previous sections each value that belongs to a cluster should not be treated equally. Thus, in order to handle uncertainty, we define mapping functions for the clusters, based on fuzzy logic. These functions map the clusters to the fuzzy domain and enable the production of classification uncertainty during the classification process. Then, we use the clusters' definition (e.g. representatives of clusters) and membership functions, in order to define the classification approach introduced in this paper.

The basic modules of the system follow (Figure 1):

- *Definition of initial classes:* A clustering or classification algorithm discovers classes (or clusters) that correspond to the distribution of the data. The result of the module is for each non-categorical attribute (A_i) of a data set a set of classes $L=\{l_i\}$ (where l_i a category) and a set of mapping functions appropriately chosen.
- *Mapping to the fuzzy domain:* The result of this procedure is a set of degrees of belief (d.o.bs) $\{M = \{\mu_{l_i}(t_k.A_i)\}$. 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 . The resulting d.o.bs are stored in a structure called Classification Value Space (CVS).
- *Quality Assessment:* In this module the quality of the classification scheme is assessed in terms of information measures extracted from the CVS. The goal is to assess how well the current classification model is applied to the data set. As the database grows and new instances are classified the system is able to check if it is necessary to redefine the initial clustering scheme.
- *Queries and Decision support.* The CVS includes significant knowledge for our data set. We can exploit this knowledge for decision making, based on the energy metric [5] measure. Then, we exploit the results of these measures in order to make decisions with reference to the knowledge conveyed by CVS.

4 Mapping to the fuzzy domain

In this section, we briefly present the procedures for uncertainty representation after the definition of the initial categories on which the classification process is based. The interested reader may consult [19] for more details.

4.1 Classification space (CS)

Assuming a data set S , we define the initial groups into which our data can be partitioned. A clustering algorithm can be used in order to identify these initial groups of data (clusters), which then used in the classification process. As mentioned before, there is inherent uncertainty in the classification of a value in a set of classes. The set of clusters can be represented by the clusters' representatives and the extent of the

partitions. Then, we attach to each class a mapping function that maps each real value to the fuzzy domains and represents the belief that the value belongs to the class. We introduce the term *Classification Space* (CS) that implies the specifications of the classes along with the attached mapping functions. Assuming the appropriate set of value domains for these classes, for each attribute (or set of attributes) 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 user defined lexical category that corresponds to cluster i , $[v_1, v_2]$ is the respective value interval and f_i the assigned mapping function. The value domains may be overlapping, increasing thus the expressive power of the classification mechanism since some values may be classified to one or more classes with different d.o.bs.

The selection of mapping functions is an important issue that can affect the results of the classification process. However, in this paper we do not deal with the evaluation of membership functions or their influence to the classification results. We have currently adopted the *hypertrapezoidal membership functions (HMFs)* [8] (see Figure 2), though we can use any other type of membership functions [5]. The main reason of HMFs selection is that they are proposed as a convenient mechanism for representing and dealing with multidimensional fuzzy sets. The definition of these functions is based on the representatives of clusters and a factor, which determines the ambiguity (overlapping) between the clusters [8]. Thus, we can use them as the appropriate functions for representing the uncertainty in multidimensional datasets.

4.2 Classification Value Space (CVS)

The result of mapping the data set values to the fuzzy domain using the CS can be represented by a 3D structure, further called Classification Value Space (CVS) (see Figure 3). 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. $\mu_{ij}(S.t_k.A_i)$. In the sequel, we refer to 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.

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

```

for each attribute  $A_i$  in CS
  for each category  $l_j$  of  $A_i$ 
    for each value  $t_k.A_i$  in the data set
      compute  $\mu_{1i}(S.t_k.A_i)$ 
    end
  end
end

```

The time complexity of the above algorithm is $O(d \cdot c \cdot n \cdot f(c, d))$ where d is the number of attributes (data set dimension), c is the number of classes (clusters), n is the number of d.o.b. values for a category l_j (number of tuples in the data set) and $f(c, d)$ is the complexity of degrees of belief computation. Usually $c, d \ll n$. Thus, the time complexity for computing the d.o.bs for a data set will be $O(n)$.

4.3 CVS Storage

As the expected size of the CVS is very big we designed a scheme that minimizes the storage requirements. The CVS is stored in two database tables, further called *cube dictionary* and *CVS table* respectively. The *cube dictionary* includes information about the CS of the data sets. It stores the name of cube, the name of the data set (i.e., database table) that a specific cube refers to, the name of the data set attributes, the names of the data set classes and the membership function assigned to each of them. The *CVS table* corresponds to CVS cube. It stores the distinct values of our data set and the degree of

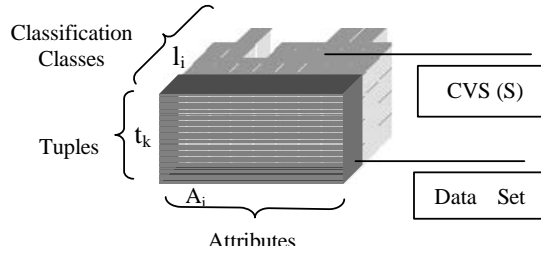


Figure 3. The CVS holding the “degrees of belief” (d.o.b.s) for the classification of the attributes’ values

belief with which a specific value is classified into pre-specified classes of our data set. Each category (cluster) of a data set corresponds to a specific column of the table while there is also a specific column for each attribute of the data set. Each distinct value of the data set and the corresponding d.o.b. is stored only once. Thus, there are no duplicate values and the storage requirements of the proposed structure (i.e., CVS) are minimized to the storage cost for the data set distinct values. Assume a data set S with N tuples and let $\text{Dist}(N)$ the number of distinct values of S . Then the cost of storage for cube related to S will be $\text{Dist}(N)/N$ of the cost for the whole data set. For instance, in case of a corporate data set consisting of 1000 tuples, there are 74 different values for the attribute “age”. Thus, the storage cost of the cube related to “age” classification will be 0.07 of the cost for the whole data set.

5 Information Measures for decision support

The CVS conveys significant knowledge included in cumulative information measures. Various information measures have been proposed in literature such as entropy and energy [5]. We adopt the energy metric, which is essentially a measure of the overall information content of a fuzzy set (in our case CVS). We exploit it in order to evaluate classification schemes or support decision making related to a data set. The aim here is to be able to compare

- i. relative importance of classes in a data set (e.g., “young vs. old customers”)
- ii. relative importance of classes across data sets
- iii. the information content of different data sets

In the sequel, we assume both the case of one- and multi- dimensional classification. This means that we can define classes for our data set and the corresponding membership functions of its initial classes, taking in account one or more attributes (e.g. “salary”, “age”, “salary and age”).

5.1 Class energy metric.

This is a measure of the information (significance) of a class l_i in the data set S . Let A_i be a set of attributes ($A_{i1}, A_{i2}, \dots, A_{im}$) and l_i a related category. Then the overall information that S contains, regarding the classification of its data in the category l_i is given by the information measure:

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

where q is a positive integer. The typical value of q is 2. Higher values suppress lower d.o.b. making the contribution of the tuples with high (close to 1) d.o.bs more significant.

For instance assume an attribute “salary” and its category *high*. Applying Equation 1 the overall information included in the data set for the category *high* salaries is given by

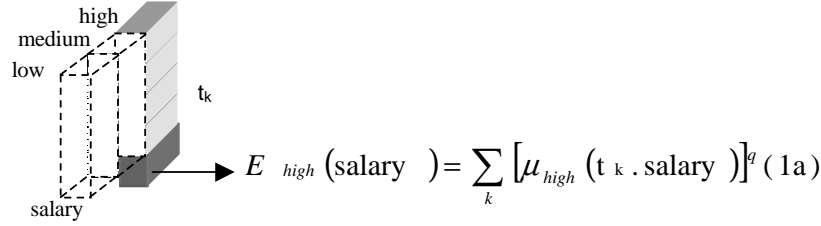


Figure 4. Representing the *category energy metric* in cube.

the formula 1a. In Figure 4 the corresponding slice and column of CVS are selected in order to acquire the information measure E_{high} .

5.2 Attribute energy metric.

The *overall energy* of a set of attributes $A_i = (A_{i1}, A_{i2}, \dots, A_{im})$, is the sum of the energy metric values for all the attribute classes. This measure represents the information content of the attribute. Hence:

$$E_{A_i}(S) = \sum_{li} E_{li} \quad (2)$$

More specifically, $E_{A_i}(S)$ represents the information included in the slice of the CVS cube (Figure 5) corresponding to a specific attribute. For instance, the slice of cube in Figure 5 represents the overall information for attribute “salary” when we classify the data in the classes *low*, *medium* and *high*.

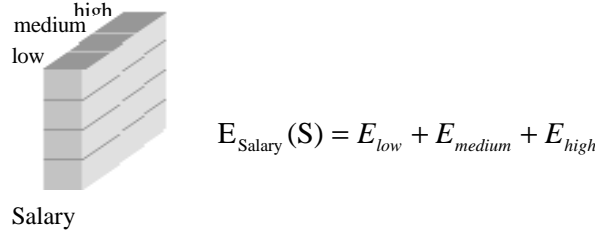


Figure 5. Attribute energy metric in the CVS

6 Classification scheme quality assessment

One of the most important requirements is the assessment of the classification scheme’s quality. This implies how successful a classification scheme is, considering a specific data set and how well the defined classes of an attribute fit the data.

A successful classification scheme should contain a significant amount of information i.e., the value of class/attribute *energy* have to be as high as possible. Another requirement is the minimization of the entropy in the defined classes, i.e., to minimize the cases that the data values are equally assigned to all classes. We introduce a new quality assessment index for classification based on these criteria and concepts of the information theory.

Let $C = \{c_1, \dots, c_{n_c}\}$ to be a classification scheme for a data set S into n_c clusters. The following measures are defined to assess the quality of a classification scheme.

Uncertainty of a class. It evaluates the uncertainty within a class based on the memberships (degrees of belief) of the data into the specific class. This term is also known in the information theory as *surprise*[20]:

$$Unc_Cl_{c_j} = - \sum_{i=1}^N \log_2(\mu_{ij}) / N \quad (3)$$

where N is the number of tuples in the dataset under consideration

Overall belief of a class. The *overall belief* that a data set supports a set of classes is given by the equation:

$$DoB_{c_j} = \sum_{i=1}^N \mu_{ij}^q / N \quad (4)$$

where N is the number of tuples in the data set S. This is also the value of $E_{A_i}(S)$ in case of crisp classification (i.e., each value belong to one and only one category). In case that all membership values to a class are equal DoB_{c_j} obtain its higher value, i.e., $\log_2(n_c)$. Moreover, this is an indication that the class c_j does not fit the data under consideration.

Information coefficient of a class. It is an index of the quality of the class under consideration defined as

$$InfoCl_i = DoB_{c_j} \cdot (\log_2(n_c) - Unc_Cl_i) \quad (5)$$

where n_c is the number of classes under consideration.

The definition of *Info_Coef* indicates that both criteria of a “good” classification scheme (i.e., amount of information and uncertainty) are properly combined, enabling reliable evaluation of results. The first term, DoB_{c_j} , indicates the significance of a class in the data set, i.e., the amount of information included in the specific class. A high value of this term is an indication of a class that is significantly supported by the data. The second term is an indication of the class uncertainty. More specifically, it evaluates the deviation of the class uncertainty from the case that all membership values to a class are equal (i.e., the case of no clustering tendency or improper definition of classes). The highest is the value of this term the highest is our belief that the data are classified to the proper class and thus the defined scheme fits to the data set under consideration. Then the *Information coefficient of the classification scheme C* is given by the equation:

$$Info_Coef(C) = \frac{1}{n_c} \sum_{i=1}^{n_c} InfoCl_i \quad (6)$$

where n_c is the number of classes under consideration.

Thus, the *Info_Coef* can be used as a measure for finding the best partitioning that fits a data set taking in account the uncertainty included in its values. We consider a variety of classification schemes for our data set, as defining by considering the results of different clustering algorithms. Then, we evaluate them based on the *Info_Coef* measure in order to select the scheme that best fit our data set. In general terms, the best classification scheme corresponds to a local maximum of *Info_Coef* in its graph versus n_c (number of clusters/classes). It is the point (here the number of classes) at which the *Info_Coef* is maximized.

7 Experimental Study

Based on the framework we described in previous sections, we implemented a classification system for handling uncertainty in the data mining process. It is a system implemented in Java using the JDBC application interface to connect to a data set. Using this system, we experimented with synthetic data sets of known structure. The experimental results focus on handling the classification belief and exploiting it for decision-making. Also, we evaluate the use of information measures for assessing the quality of classification schemes.

Table 1. (a) Category and Energy metrics for “salary”, (b) Category and Energy metrics for “age”

<i>Salary</i>	
E_{low}	304.2736
E_{medium}	343.2594
E_{high}	348.0000
E_{salary}	995.5320

<i>Age</i>	
E_{young}	398.75
E_{old}	540.96
E_{age}	939.71

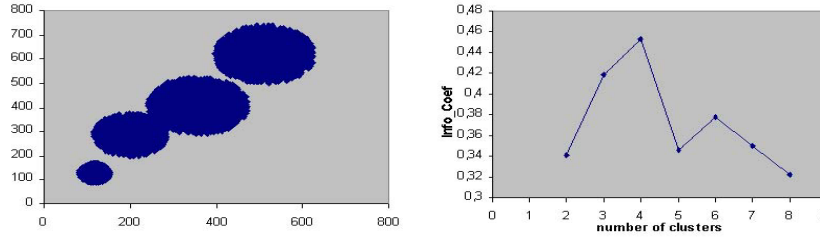


Figure 6. **a.** A data set classified in four clusters, **b.** The graph of QoC_{Ai} versus the number of clusters considering a synthetic two-dimensional data set

We assume a synthetic data set maintaining information related to the employees. The schema of this data set is $R = \{\text{salary, age}\}$. Our system uses the CS (centers, name of categories, value domain and mapping functions) for the data set and transforms it into a CVS. This implies classification of the data set values into classes using HMFs as the mapping functions. Table 1a and Table 1b present the class energy metric values for the attributes “salary”, “age” respectively. Some “useful” knowledge about data set can be extracted from these tables. For instance, the class energies in Table1a indicate that our data set supports with more confidence *high* salaries than *low* salaries ($E_{high} > E_{low}$). Also, Table1b indicates that in this data set we are more confident to have *old* employees than *young* ones ($E_{old} > E_{young}$).

Selecting the optimal classification scheme. This part of our experiments refers to quality assessment of classification schemes. We experiment with real and synthetic data sets and the goal is to evaluate the different classification schemes, resulting from different learning procedures, so as to select the scheme that best fits our data set.

The evaluation of schemes is based on the quality classification measure $Info_Coef$ defined in Section 6. More specifically, we considered different partitionings of a data set corresponding to the different sets of initial classes on which the classification process will be based. Then exploiting the appropriate membership functions (in our case HMFs) we map the values to the fuzzy domains. Thus, the data set is transformed into CVSs, one for each classification scheme (partitioning). Using the $Info_Coef_{Ai}$ measure we evaluate the defined classification schemes so as to select the scheme that best fits the data under consideration. In the sequel, due to lack of space, we present only some representative examples of our experimental study.

We consider a synthetic two-dimensional data set, following the normal distribution. It is clear from Figure 6a that the data set consists of four overlapping clusters. This is also verified by our approach based on the $Info_Coef$ measure. Figure 6b depicts the behavior of $Info_Coef$ versus the number of classes. We observe there, that the classification scheme of four classes corresponds to a maximum value of $Info_Coef$. This is an indication that the best classification for the data under consideration is the scheme of four classes.

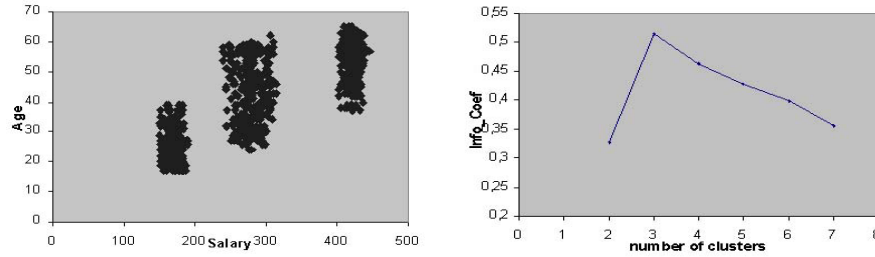


Figure 8. a. A data set classified in three classes, b. The graph of QoC versus the number of clusters considering a two-dimensional data set “salary and age”.

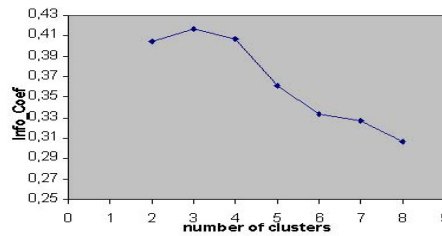


Figure 7. The graph of QoC_{A_i} versus the number of clusters considering Iris Data Set

Also, we use a sample data set (see Figure 8a) that contains the values of “salary and age”. We apply two-dimensional clustering so as to define the initial classes of the data set considering the attributes “salary and age”. Figure 8b shows the graph of the *Info_Coef* measure as a function of the number of clusters. We observe there, that the classification scheme with three classes corresponds to a maximum value of *Info_Coef*. This is an indication that the best classification for “salary and age” a scheme of three classes, which is also verified by the distribution of the data set values in Figure 8a.

We carried out a similar experiment using Iris Data Set. It consists of 150 measurements (length and width of sepal and petal) belonging to three flower varieties. This is also verified by our approach. We consider eight different classifications of Iris Data and we evaluate them based on the *Info_Coef* measure. As Figure 7 depicts *Info_Coef* takes its maximum value when we consider a scheme of three classes. This is an indication that the classification scheme of three classes is the scheme that best fits Iris Data.

8 Conclusion and further work

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*. This will be achieved by classifying the data into classes that fit the data set properties to a satisfactory degree. The contributions of this paper are summarized as follows:

- *Maintenance of classification belief* all the way through the classification process. A data set value can be assigned to more than one classes with a different belief.
- *Information measures* enabling decisions related to: i. relative importance of classes in a data set (i.e., “young vs. old customers”), ii. relative importance of classes across data sets, iii. the information content of different data sets
- *Quality assessment of classification models*, so as to find how well a model fits the underlying data set.

It is important to stress that our contribution is independent of the technique used for the definition of initial clusters. Indeed, we take as input the classes resulting from the

application of any algorithm on a training set and we classify all the data set to these classes introducing uncertainty features. Moreover we take into account aggregate beliefs that will assist for decision support in the data set and across data sets.

Further work in the incremental production of optimal classification and association rules extraction models. We aim at exploiting the classification quality measure presented in this paper so as to define a procedure for evaluating classification models through out the life cycle of a data set as insertions/updates and deletions occur. Also, different mapping functions and their effect to the proposed classification scheme as regards uncertainty representation will be studied. Moreover, alternative information measures proposed in literature will be tested and 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 Wiley & Sons, Inc, 1996.
- 2.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.
- 3.P. Cheeseman, J. Stutz. "Bayesian Classification (AutoClass): Theory and Results". *Advances in Knowledge Discovery and Data Mining*. (Eds:U. Fayyad,et al), AAAI Press,1996.
- 4.U. Fayyad, G. Piatetsky-Shapiro, P. Smuth & R. Uthurusamy(editors). "From DataMining to Knowledge Discovery: An Overview". *Advances in Knowledge Discovery and Data Mining*. AAAI Press, 1996.
- 5.M. Gupta, and T. Yamakawa, (eds). "Fuzzy Logic and Knowledge Based Systems", *Decision and Control* (North Holland). 1988.
- 6.M. Halkidi, M. Vazirgiannis. Clustering: Quality measures and uncertainty handling. *Technical report*, Athens Univ. of Economic & Business, 1999
- 7.T. Horiuchi. "Decision Rule for Pattern Classification by Integrating Interval Feature Values". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.20, No.4, April 1998, pp.440-448.
- 8.W. Kelly, J. Painter. "Hypertrazoidal Membership Functions". *5th IEEE International Conference on Fuzzy Systems*, New Orleans, September 8, 1996.
- 9.M. Melta, R. Agrawal, J. Rissanen. "SLIQ: A fast scalable classifier for data mining". In *EDBT'96, Avigon France*, March 1996.
10. T. Mitchell. *Machine Learning*. McGraw-Hill, 1997
11. J.R Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufman, 1993.
12. R. Rastori, K. Shim. "PUBLIC: A Decision Tree Classifier that Integrates Building and Pruning". *Proceeding of the 24th VLDB Conference*, New York, USA, 1998.
13. J.Shafer, R. Agrawal, M. Mehta. "SPRINT: A scalable parallel classifier for data mining". In *Proc. of the VLDB Conference*, Bombay, India, September 1996
14. Glymour C., MadiganD., Pregibon D, Smyth P, "Statistical Inference and Data Mining", in *CACM* v39 (11), 1996, pp. 35-42
15. Cezary Z. Janikow, "Fuzzy Decision Trees: Issues and Methods", *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 28, Issue 1, pp 1-14, 1998.
16. M. Vazirgiannis, "A classification and relationship extraction scheme for relational databases based on fuzzy logic", in the proceedings of the Pacific -Asian Knowledge Discovery & Data Mining '98 Conference, Melbourne, Australia, 1999.
17. S. Theodoridis, K. Koutroubas. *Pattern recognition*, Academic Press, 1999
18. Bezdeck J.C, Ehrlich R., Full W., "FCM:Fuzzy C-Means Algorithm", *Computers and Geoscience* 1984
19. M. Vazirgiannis, M. Halkidi. "Uncertainty handling in the datamining process with fuzzy logic", to appear in the proceedings of the IEEE-FUZZ conference, San Antonio, May, 2000.
20. T. Shneider. "Information Theory Primer", Chapter II, PhD thesis: "The information Content of Binding Sites on Nucleotide Sequences". <http://www.lecb.ncifcrf.gov/~toms/paper/primer/>