

UMiner: A Data mining system handling uncertainty and quality

Christos Amanatidis, Maria Halkidi, Michalis Vazirgiannis
Dept of Informatics, Athens Univ. of Economics and Business
{krisaman, mhalk, mvazirg}@aueb.gr

Abstract In this paper we present UMiner, a new data mining system, which improves the quality of the data analysis results, handles uncertainty in the clustering & classification process and improves reasoning and decision-making.

1 Introduction

The explosive growth of data collections in the science and business applications and the need to analyse and extract useful knowledge from this data leads to a new generation of tools and techniques grouped under the term data mining [3]. Their objective is to deal with volumes of data and automate the data mining and knowledge discovery from large data repositories. Most data mining systems produce a particular enumeration of patterns over data sets accomplishing a limited set of tasks, such as clustering, classification and rules extraction [1, 2]. However, there are some aspects that are under-addressed by the current approaches in database and data mining applications. These aspects are: i) *the reveal and handling of uncertainty* in context of data mining tasks. ii) *the evaluation of data mining results* based on well established quality criteria. In this paper we present a data mining framework to evaluate the data analysis results, to handle efficiently the uncertainty in the context of classification process and exploit the classification belief in the process of reasoning and decision-making. Then, we present UMiner, a client/server system that we have developed based on this framework while we describe their architecture and its main services.

2 UMiner development approach

The importance of the requirements discussed above in the data mining process, that is the usage and reveal of uncertainty and the evaluation of data mining results, led us to the development of UMiner. Fig. 1 depicts the overall framework on which UMiner's architecture is based. The major tasks of the system can be summarised as follows:

- *Clustering*. In this step we define/extract clusters that correspond to the initial categories for a particular data set. We can use any of the well-established clustering methods that are available in literature.
- *Evaluation of the clustering scheme*. The clustering methods can find a partitioning of our data set, based on certain assumptions. Thus, an algorithm may result in different clustering schemes for a data set assuming different parameter values. UMiner evaluates the results of clustering algorithms based on a well-defined quality index [6] and selects the clustering scheme that best fits the considered data. The definition of this index is based on the two fundamental criteria of clustering quality, which are compactness and well separation.
- *Definition of membership functions*. Most of the clustering methods do not handle uncertainty i.e., all values in a cluster belong totally to it. We introduce a new approach so as to transform crisp clustering schemes into fuzzy ones. This is achieved by defining a scheme for assignment of the appropriate membership functions to the clusters. In the current system the implementation of membership function is based on the *Hypertrapezoidal Fuzzy Membership Functions*[4]. Moreover, a number of some other well-known membership functions are supported such as triangular, linear decreasing or linear increasing functions.

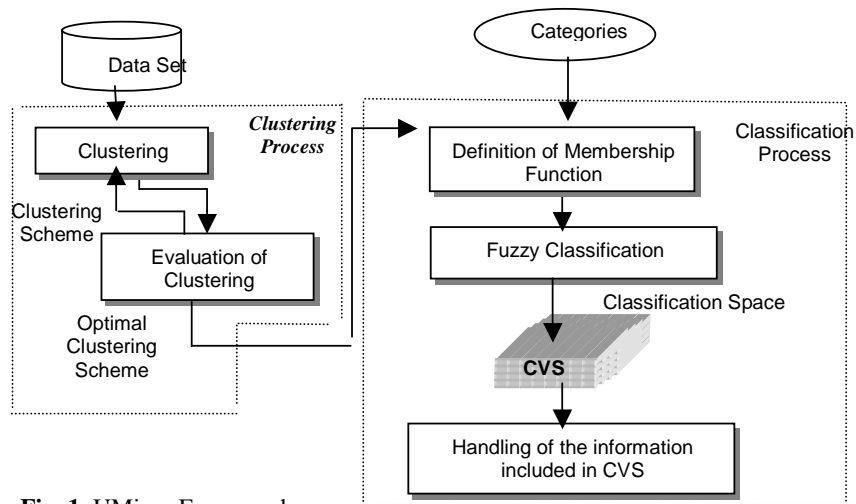


Fig. 1. UMiner Framework

- *Fuzzy Classification.* The values of the non-categorical attributes (A_i) of a data set are classified into categories according to a set of categories $L=\{l_i\}$ (where l_i a category) and a set of classification functions defined in preceding clustering process.
- *Classification Value Space (CVS) Construction.* According to the classification framework proposed in [7] we transform the data set into classification beliefs and store them into a structure called CVS. It is represented by a cube, the cells of which store the degrees of belief for the classification of the attributes' values.
- *Handling of the information included in the CVS.* The CVS includes significant knowledge for our data set. We can exploit this knowledge for decision-making, based on well-established information measures that we define in [7].

3 System Architecture - Demonstration

UMiner is a data mining client-server system based on the framework described in Sect 2. The server provides the major services of system: i) access to data, ii) implementation of clustering algorithms, iii) CVS construction. The connection to the declared databases is performed through the JDBC application interface. The client is a Java application that retrieves data and requests data mining tasks from the system server. Using the appropriate plug in it could run through a Web Browser. The major client functions are: i) authentication and connection to the server, ii) clustering and cluster validity, iii) CVS construction process, iv) visualization of data and data mining results.

We will demonstrate UMiner emphasizing its advantages. We focus on clustering quality assessment, uncertainty handling and visualization of data mining results. Following are some highlights of the system demonstration.

- **Server Connection:** After logging into the UMiner server, a list of available databases is presented to the user who may select one of them to work with. Once the user has opened a database connection, then a list of all available tables, and CVSs is presented.
- **Visualization:** Selecting a table user may view the data, using one of the available visualization techniques that our system supports. These techniques are: i) 3D, ii) matrix scatterplot, iii) glyph, iv) parallel coordinators, v) table.
- **Clustering:** The user may select one of the available clustering algorithms (i.e., K-means, DBSCAN, CURE) in order to define a partitioning for the data set. Depending on the clustering algorithm, the user defines the values of its input parameters. Then,

the algorithm partition the dataset to a specific set of clusters and the defined clusters can be presented to the end user, using one of the available visualization techniques. Each cluster is rendered with a different colour.

- **Cluster Validity:** Selecting the validation task the system searches for the optimal parameters' values for a specific clustering algorithm so as to result in a clustering scheme that best fits our data. The user selects the clustering algorithm and the input parameter based on which the validation task will be performed. Also, the range of input parameters values is defined. Then, the system presents the graph of the quality index versus the number of clusters, based on which we find the number of clusters that best fits the data set.
- **Cube Construction – Classification:** The cube construction choice gives the chance to the user to select a data set and transforms it into a CVS. The system asks the user to define the appropriate parameters, which are the attribute-categories, value domains and transformation functions. For defining categories of the data set the user has the following choices: i) to use the results of clustering produced in the previous step in order to define these parameters, ii) to give his/her own values, iii) to use the default values proposed by the system. Using the above information a set of tables is created which represents CVS. It maintains the overall classification information of the data set (i.e., the classification belief for each value in the data set).
- **CVS Information measures:** The user may select an already constructed CVS and ask for energy metrics related to it. The system computes the category energy metric and the overall energy of each attribute and presents the results to the user. Based on these results, the users can extract useful knowledge related to the data sets such as 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 quality of a classification model i.e., how well it fits a data set.

4 Conclusion

UMiner aims at supporting data mining tasks while it enhances validity and handling of uncertainty. At the current stage of development we present UMiner as a clustering and classification system that is suitably extended to support uncertainty. It is a client-server corba-based data mining system and uses the JDBC application interface to connect to a data set.

Our further work will be concentrated on the development of new modules for our system such as rules extraction, reasoning with information measures based on user queries.

References

1. M.Berry, G. Linoff. *Data Mining Techniques For marketing, Sales and Customer Support*. John Willey & Sons, Inc, 1996.
2. U. Fayyad, G. Piatetsky-Shapiro, P. Smuth and Ramasamy Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. AAAI Press 1996
3. U. Fayyad, R. Uthurusamy. “Data Mining and Knowledge Discovery in Databases”, *Communications of the ACM*. Vol.39, No11, November 1996.
4. W. Kelly, J. Painter. “Hypertrapezoidal Fuzzy Membership Functions. 5th *IEEE Int. Conf. on Fuzzy Systems*, New Orleans, Sept. 8, pp1279-1284, 1996.
5. S. Theodoridis, K. Koutroubas. *Pattern recognition*, Academic Press, 1999
6. M. Halkidi, M. Vazirgiannis, Y. Batistakis. “Quality scheme assessment in the clustering process”, *In Proceedings of PKDD*, Lyon, France, 2000.
7. M. Vazirgiannis, M. Halkidi. “Uncertainty handling in the datamining process with fuzzy logic”, in the proceedings of *the IEEE-FUZZY Conf*, Texas, May, 2000.