International Journal of Enhanced Research in Management & Computer Applications
ISSN: 2319-7471, Vol. 4 Issue 12, December-2015

# Structured Data Mining and Data Analysis using Microarray Technology

Laxmi Devi

**ABSTRACT**

**Since web-based educational systems are capable of collecting vast amounts of student profile data, data mining and knowledge discovery techniques can be applied to find interesting relationships between attributes of students, assessments, and the solution strategies adopted by students.**

**The focus of this paper is on the structured data mining and microarray data analysis. The growth of the use of semi-structured data has created new opportunities for data mining, which has traditionally been concerned with tabular data sets, reflecting the strong association between data mining and relational databases. Much of the world's interesting and mineable data does not easily fold into relational databases, though a generation of software engineers have been trained to believe this was the only way to handle data, and data mining algorithms have generally been developed only to cope with tabular data.**

**Keywords: structured, data mining, microarray, web, clustering, methodology.**

## INTRODUCTION

The ever-increasing progress of network-distributed computing and particularly the rapid expansion of the web have had a broad impact on society in a relatively short period of time. Education is on the brink of a new era based on these changes. Online delivery of educational instruction provides the opportunity to bring colleges and universities new energy, students, and revenues. Many leading educational institutions are working to establish an online teaching and learning presence. Several different approaches have been developed to deliver online education in an academic setting.

### Data Mining

Data Mining is the process of analyzing data from different perspectives and summarizing the results as useful information. It has been defined as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

Presently, the amount of data stored in databases is increasing at a tremendous speed. This gives rise to a need for new techniques and tools to aid humans in automatically and intelligently analyzing huge data sets to gather useful information. This growing need gives birth to a new research field called Knowledge Discovery in Databases (KDD) or Data Mining, which has attracted attention from researchers in many different fields including database design, statistics, pattern recognition, machine learning, and data visualization. In this chapter we give a definition of KDD and Data Mining, describing its tasks, methods, and applications. Our motivation in this study is gaining the best technique for extracting useful information from large amounts of data in an online educational system, in general, and from the LON-CAPA system, in particular. The goals for this study are: to obtain an optimal predictive model for students within such systems, help students use the learning resources better, based on the usage of the resource by other students in their groups, help instructors design their curricula more effectively, and provide the information that can be usefully applied by instructors to increase student learning.

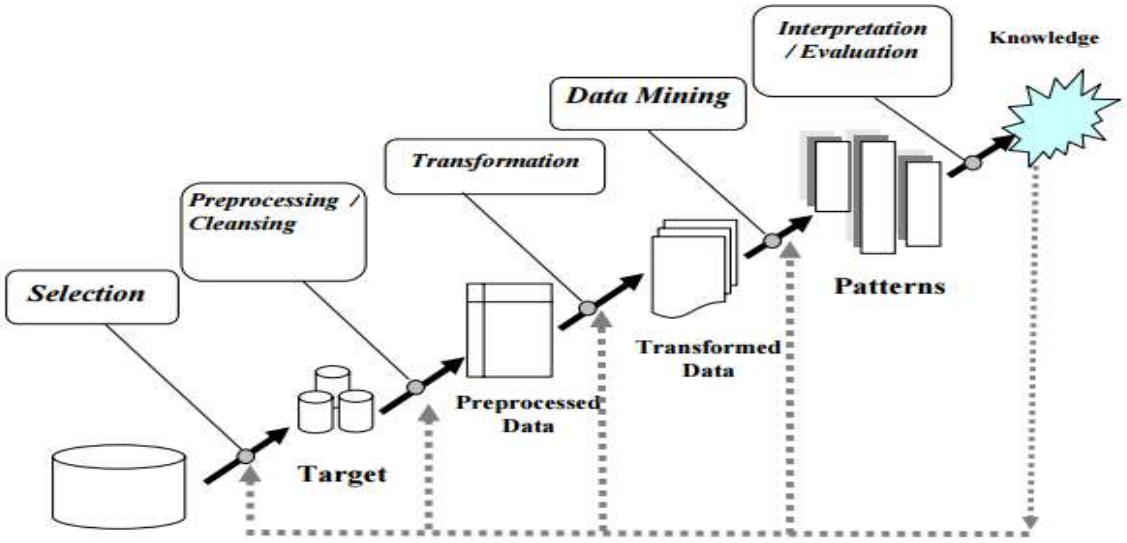


**Figure 1: Steps of the KDD Process**

The process of data mining uses machine learning, statistics, and visualization techniques to discover and present knowledge in a form that is easily comprehensible. The word "Knowledge" in KDD refers to the discovery of patterns which are extracted from the processed data.

**Structured Data Mining**

In this section, the author provides an overview of the genus of Data Mining paradigms that deal with structured data. She starts with definitions of the major concepts in Structured Data Mining having to do with how structured data is represented and how knowledge can be extracted from this data. These concepts are shared by the different SDM paradigms that are the subject of the remaining sections of this chapter. She briefly outlines each approach, and describes how they implement SDM. Subsequently, we describe on a more detailed level the strengths and weaknesses of each paradigm.

**Structure mining** or **structured data mining** is the process of finding and extracting useful information from semi structured data sets. Graph mining is a special case of structured data mining. XML, being the most frequent way of representing semi-structured data, is able to represent both tabular data and arbitrary trees. Any particular representation of data to be exchanged between two applications in XML is normally described by a schema often written in XSD. Practical examples of such schemata, for instance NewsML, are normally very sophisticated, containing multiple optional subtrees, used for representing special case data. Frequently around 90% of a schema is concerned with the definition of these optional data items and sub-trees.

Messages and data, therefore, that are transmitted or encoded using XML and that conform to the same schema are liable to contain very different data depending on what is being transmitted. Such data presents large problems for conventional data mining. Two messages that conform to the same schema may have little data in common. Building a training set from such data means that if one were to try to format it as tabular data for conventional data mining, large sections of the tables would or could be empty. There is a tacit assumption made in the design of most data mining algorithms that the data presented will be complete. The other necessity is that the actual mining algorithms employed, whether supervised or unsupervised, must be able to handle sparse data. Namely, machine learning algorithms perform badly with incomplete data sets where only part of the information is supplied.

For instance methods based on neural networks.[citation needed] or Ross Quinlan's ID3 algorithm.[citation needed] are highly accurate with good and representative samples of the problem, but perform badly with biased data. Most of times, better model presentation with more careful and unbiased representation of input and output is enough. A particularly

relevant area where finding the appropriate structure and model is the key issue is text mining. XPath is the standard mechanism used to refer to nodes and data items within XML. It has similarities to standard techniques for navigating directory hierarchies used in operating systems user interfaces. To data and structure mine XML data of any form, at least two extensions are required to conventional data mining. These are the ability to associate an XPath statement with any data pattern and sub statements with each data node in the data pattern, and the ability to mine the presence and count of any node or set of nodes within the document.

As an example, if one were to represent a family tree in XML, using these extensions one could create a data set containing all the individuals in the tree, data items such as name and age at death, and counts of related nodes, such as number of children. More sophisticated searches could extract data such as grandparents' lifespans etc. he addition of these data types related to the structure of a document or message facilitates structure mining.

As was outlined in the central subject of analysis is the individual. We assume in Structured Data Mining that an individual consist of parts that are somehow connected to form a structured individual. Parts can be thought of as small portions of an individual that are atomic: they exhibit no internal structure. All structural relations are between parts, rather than within parts. Parts typically have a number of attributes associated with them. They can thus be thought of as tuples that behave similar to the flat individuals that are the subject of Propositional Data Mining. In general, structured individuals will not be arbitrary collections of parts. Parts will appear in a relatively small number of types, referred to as classes. All parts, over all individuals, are instances of one of the classes. A class determines which attributes are available for all instances of that class.

Not just the characteristics of parts are important in Structured Data Mining, but also how they relate to form structured individuals. We will think of individuals as annotated graphs, where the nodes represent the parts. Labelled, but undirected, edges between parts represent the relationship between pairs of parts. Other than the label, the edge provides no information concerning the relationship between the parts. Typically, there will not be edges between arbitrary pairs of parts. Most data representation schemes will only allow relations between specific classes of parts to enforce certain types of structure in the individuals. Furthermore, there will often be restrictions on the number of parts of a certain class that may be related to a part of some other class, and vice versa. A definition of the restrictions on relations between parts of two classes will be referred to as an association between two classes. Edges in an individual can thus be thought of as instances of an association, where the label refers to the association. It should be noted that this use of the term association is not related to the term association rules, which indicates a popular family of models of statistical dependency. The present associations are hard constraints on the data.

## Data Analysis using Microarray Technology

In the last years, with the developing of new technologies and revolutionary changes in biomedicine and biotechnologies, there was an explosive growth of biological data. Genome wide expression analysis with DNA microarray technology has become a fundamental tool in genomic research. Since microarray technology was introduced, scientists started to develop informatics tools for the analysis and the information extraction from this kind of data. Due to the characteristics of microarray data (i.e. high levels of noise, high cardinality of genes, small samples size) data mining approaches became a suitable tool to perform any kind of analysis on these data. Many techniques can be applied to analyze microarray data, which can be grouped in four categories: classification, feature selection, clustering and association rules. Classification is a procedure used to predict group membership for data instances. Given a training set of samples with a specific number of attributes (or features) and a class label (e.g., a phenotype characteristic), a model of classes is created. Then, the model is exploited to assign the appropriate class label to new data. Model quality is assessed by means of the classification accuracy measure, i.e., the number of correct label predictions over the total number of unlabeled data. The classification of microarray data can be useful to predict the outcome of some diseases or discover the genetic behavior of tumors. Since genetic data are redundant and noisy, and some of them do not contain useful information for the problem, it is not suitable to apply the classification directly to the whole dataset. Feature selection techniques are dimensional reduction methods usually applied before classification in order to reduce the number of considered features, by identifying and removing the redundant and useless ones. Moreover, feature selection algorithms applied to microarray data allow identifying genes which are highly correlated with the outcome of diseases.

Analysis of microarrays presents a number of unique challenges for data mining. Typical data mining applications in domains like banking or web, have a large number of records (thousands and sometimes millions), while the number of

fields is much smaller (at most several hundred). In contrast, a typical microarray data analysis study may have only a small number of records (less than a hundred), while the number of fields, corresponding to the number of genes, is typically in thousands. Given the difficulty of collecting microarray samples, the number of samples is likely to remain small in many interesting cases. However, having so many fields relative to so few sample, creates a high likelihood of finding "false positives" that are due to chance – both in finding differentially expressed genes, and in building predictive models. We need especially robust methods to validate the models and assess their likelihood.

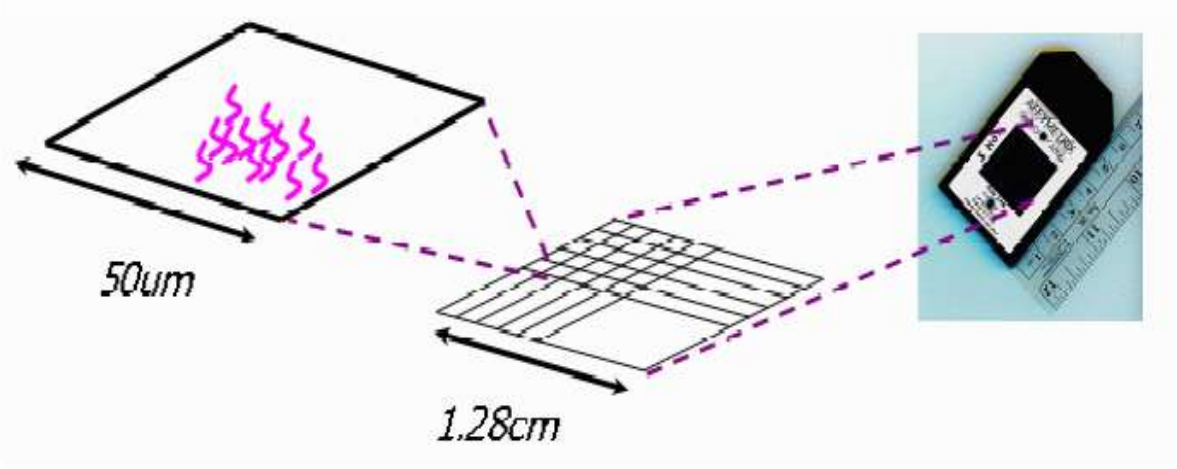


**Figure 2: Affymetrix GeneChip (right),
its grid (center) and a cell in a grid (left)**

Microarrays have opened the possibility of creating data sets of molecular information to represent many systems of biological or clinical interest. Gene expression profiles can be used as inputs to large-scale data analysis, for example, to serve as fingerprints to build more accurate molecular classification, to discover hidden taxonomies or to increase our understanding of normal and disease states. The first generation of microarray analysis methodologies developed over the last 5 years has demonstrated that expression data can be used in a variety of class discovery or class prediction biomedical problems including those relevant to tumor classification. Machine learning and statistical techniques applied to gene expression data have been used to address the questions of distinguishing tumor morphology, predicting post treatment outcome, and finding molecular markers for disease. Today the microarray-based classification of different morphologies, lineages and cell histologies can be performed successfully in many instances. The performance in predicting treatment outcome or drug response has been more limited but some of the results are quite promising. Most results of microarray analysis still require further experimental validation and follow up study.

**Methodology**

This is the procedure used in evaluating the various predictive data mining techniques using the four different and unique data sets. This section also deals with the introduction and description of the four data sets used in this study, using preliminary diagnoses to check for the relationships between the variables in each data set and to visualize the nature or properties of the data sets. Figure 3 shows a diagram of the methodology used in this work.

The four data sets are first introduced, as well as the preliminary diagnoses done on each data set to gain an insight into their properties. The relationship check is made by plotting the inputs over the output of the raw data sets. The data is preprocessed by scaling or standardizing them (data preparation) to reduce the level of dispersion between the variables in the data set. The correlation coefficients of each of the various data sets are computed to verify more on the relationship between the input variables and the output variables. This is followed by finding the singular value decomposition of the data sets transforming them into principal components. This also will be helpful in checking the relationship between the variables in each data set. At this stage, the data sets are divided into two equal parts, setting the odd number data points as the "training set" and the even number data points as the "test validation data set." Now the train data for each data set is used for the model building. For each train data set, a predictive data mining technique is used to build a model, and the various methods of that technique are employed.

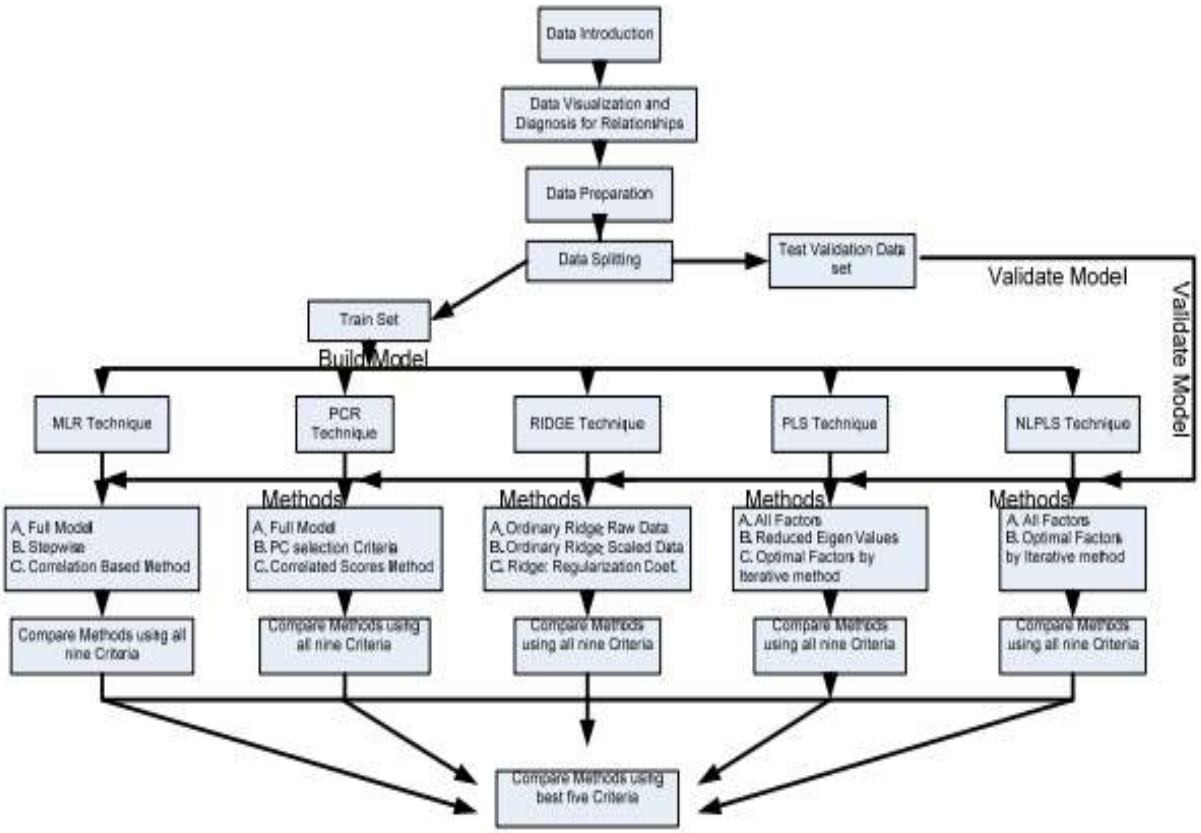


**Figure 3. Flow Diagram of the Methodology**

## CONCLUSIONS

Structured data is getting more and more important in database applications, such as molecular biology, image retrieval or XML document retrieval. Attributed graphs are a natural model for the structured data in those applications. For the clustering and classification of such structured data, a similarity measure for attributed graphs is necessary. All known similarity measures for attributed graphs are either limited to a special type of graph or computationally extremely complex, i.e. NP-complete, and are, therefore, unsuitable for data mining in large databases.

Also, Microarrays are a revolutionary new technology with great potential to provide accurate medical diagnostics, help find the right treatment and cure for many diseases and provide a detailed genome-wide molecular portrait of cellular states. The papers included in this issue are a good sample of second generation methodologies and techniques that are being used or under development today. They are very promising and extend the possibilities of applying computational analysis and data mining to aid research in biology and medicine.

## REFERENCES

[1]. AnHai Doan et. al, "The Case for a Structured Approach to Managing Unstructured Data".
[2]. B. Fortuna, M. Grobelnik, D. Mladenic, "Background Knowledge for Ontology Construction," Proc. 15th Int'l Conf. World Wide Web(WWW '06), pp. 949-950, 2006.
[3]. Cook, D.J., Holder, L.B.: Graph-based data mining. IEEE Intelligent Systems 15 (2000) 32–41.
[4]. Dr. Jagannath Aghav, Anil Vegiraju ,Harish Jadhao, "Semantic Tool for Analysing Unstructured data".
[5]. Papadopoulos, A., Manolopoulos, Y.: Structure-based similarity search with graph histograms. In: Proc. DEXA/IWOSS Int. Workshop on Similarity Search, IEEE Computer Society Press (1999) 174–178.
[6]. Iyer, V.R. et al., "The transcriptional program in the response of human fibroblasts to serum", Science 283: 83-87, (1999).

[7]. DeRisi J, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 1996 Dec;14(4):457-60.
[8]. Petrakis, E.: Design and evaluation of spatial similarity approaches for image retrieval. Image and Vision Computing 20 (2002) 59–76.
[9]. Kuhn, H.: The hungarian method for the assignment problem. Nval Research Logistics Quarterly 2 (1955) 83–97.
[10]. Munkres, J.: Algorithms for the assignment and transportation problems. Journal of the SIAM 6 (1957) 32–38.
[11]. Roussopoulos, N., Kelley, S., Vincent, F.: Nearest neighbor queries. In: Proc. ACM SIGMOD, ACM Press (1995) 71–79.
[12]. Schena, M. et al Quantitative monitoring of gene expression patterns with a cDNA microarray. Science 270:467-470 (1995).
[13]. DeRisi, J.L. et al. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278: 680-686
[14]. (1997).
[15]. Chu, S. et al., "The transcriptional program of germ cell development in budding yeast", Science 282:699-705 (1998).
[16]. Zhang, K., Statman, R., Shasha, D.: On the editing distance between unordered labeled trees. Information Processing Letters 42 (1992) 133–139.
[17]. Zhang, K., Wang, J., Shasha, D.: On the editing distance between undirected acyclic graphs. International Journal of Foundations of Computer Science 7 (1996) 43–57.