Integrated Intelligent Research (IIR)
International Journal of Computing Algorithm
Volume: 03 Issue: 03 December 2014 Pages: 260-263
ISSN: 2278-2397

An Overview of Data Mining Concepts Applied in Mathematical Techniques

G. Suresh[1], I.A.Selvakumar[2]
[1]Asst. Prof., PG & Research Department of Computer Applications,
St. Joseph's College of Arts & Science (Autonomous), Cuddalore ,Tamil Nadu, India.
[2]Research Scholar, PG & Research Department of Computer Science,
St. Joseph's College of Arts & Science (Autonomous), Cuddalore ,Tamil Nadu, India.
Email: sureshg2233@yahoo.co.in, iaselva@yahoo.com

Abstract - Data Mining and Knowledge Discovery in Databases (KDD) are rapidly evolving areas of research that contribute to several disciplines including Mathematics, Statistics, Pattern Recognition, Visualization, and Parallel Computing. This paper is projected to serve as an overview to the concepts of these rapidly evolving research and application areas. The basic concepts of these areas are outlined in this paper with some key ideas and motivate the importance of data mining concepts applied in Mathematical Techniques.

Keywords: Data Mining, Knowledge Discovery in Database, Pattern Recognition, Visualization, Parallel Computing.

I. INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. The techniques in data mining allow users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining is the search for new valuable and nontrivial information from large volumes of data which discovers an interesting outcome by discovering through either automatic or manual methods. Best results are achieved by balancing the knowledge of human experts in describing the problems and goals with the search capabilities of computers. This paper presents a comprehensive study on Data mining concepts applied in Mathematical Techniques. Section II provides a brief overview of Data mining concepts applied in various mathematical techniques. Finally, we summarize and conclude the study in Section III.

II. OVERVIEW OF DATA MINING CONCEPTS

Data mining is one of the fastest growing fields in the computer industry. Once a small interest area within computer science and statistics, it has quickly expanded into a field of its own. One of the greatest strengths of data mining is reflected in its wide range of methodologies and techniques that can be applied to the problem sets. Since data mining is a natural activity to be performed on large data sets, it has a largest target markets in the entire data warehousing, data-mart and decision support system which encircling professionals from industries such as retail, manufacturing, telecommunications, healthcare, insurance and transportation. In the business community, data mining can be used to discover new purchasing trends, plan investment strategies, minimize expenditures in the accounting system. Data mining techniques can be applied to problems in business processes, in which the goal is to understand interactions and relationships among business practices and organizations. The law enforcement and special investigative units also uses the data mining techniques to identify the fraudulent activities and discover crime trends. Data mining techniques are used by people in intelligent community who were in the part of activities relating to matters of national security and it also employed by the people who were in commercial activities.

A. KDD and Data Mining

The term data mining is often used as a synonym for the process of extracting useful information from databases. This paper draws a distinction between "Knowledge Discovery in Databases", which we call KDD, and "data mining". The term data mining has been mostly used by statisticians, data analysts, and the database communities. The earliest uses of the term come from statistics and its usage in most settings was negative with meanings of blind exploration of data without prior hypotheses to be verified. However, notable exceptions can be found. For example, as early as 1978[1], the term is used in a positive sense in a demonstration of how generalized linear regression can be used to solve problems that are very difficult for humans and the traditional statistical techniques of that time to solve. The term KDD was coined at the first KDD workshop in 1989 [2] to emphasize that "knowledge" is the end product of a data-driven process.

B. Basic Definitions:

The definitions of KDD and data mining provided in [3] as follows:

Knowledge Discovery in Databases: is the process of identifying valid, novel, potentially useful, andultimately understandable structure in data. This process involves selecting or sampling data from a data ware-house, cleaning or preprocessing it, transforming or reducing it if needed, applying a data mining component to produce models or patterns, and then evaluating the derived models or patterns. A pattern is defined to be a thrifty description of a subset of data. A model is typically a description of the entire data.

Data Mining: is a step in the KDD process concerned with the algorithmic means by which patterns ormodels are enumerated from the data under acceptable computational efficiency limitations.

The two primary goals of data mining are prediction and

International Journal of Computing Algorithm
Volume: 03 Issue: 03 December 2014 Pages: 260-263
ISSN: 2278-2397

description. Prediction involves using some variables are fields in the data set to predict unknown or future values. Description focuses on finding patterns describing the data that can be interpreted by humans. Therefore data mining activities can be put into one of the two categories:

- *Predictive data mining:* It produces the model of the system described by the given data set.
- *Descriptive data mining:* It produces new, nontrivial information based on the available data set.

In predictive data mining the goal is to produce a model, expressed in executable code and which can be used to perform classification, prediction, estimation or other similar tasks. The descriptive data mining the goal is to gain an understanding by analyzing the patters and relationships in large data sets. The goals of prediction and description can be achieved by using the following data mining concepts:

1. *Classification* – it is a predictive learning function that classifies a data item into one of several predefined classes.
2. *Regression* - it is a predictive learning function which maps a data item to a real-value prediction variable.
3. *Clustering* – it is a common descriptive task which groups the data records into subsets where items in each subset are more similar to each other than to items in other subsets.
4. *Dependency Modeling* – finding a local model that describes significant dependencies between variables or between the values of a data set or in a part of a data set.
5. *Data Summarization* – it is a descriptive task which targets to find interesting summaries from parts of the data. For example, finding similarity between a few attributes in a subset of the data.
6. *Change and Deviation Detection* – it is the process of discovering the most significant changes in the data set.

1. *Classification*

In classification the basic goal is to predict the most likely state of a categorical variable (the class) given the values of other variables. This is fundamentally a density estimation problem. If one could estimate the probability that the class (value of Y), given the value of $x \in X$, then one could derive this probability from the joint density on Y and X . However, this joint density is rarely known and difficult to estimate. Hence one has to resort to various techniques for estimating this density, including:

• Density estimation, e.g. kernel density estimators [4] or graphical representations of the joint density [5].

• Metric-space based methods: define a distance measure on data points and guess the class value based on proximity to data points in the training set. For example, the K-nearest-neighbor method [4].

• Projection into decision regions: divide the attribute space into decision regions and associate a prediction with each. For example linear discriminate analysis determines linear separators and neural networks compute non-linear decision surfaces [6]. Decision tree or rule-based classifiers make a piecewise constant approximation of the decision surface [7]. The third class of methods is the most commonly used and studied. It is usually more practical because it sidesteps the harder problem of determining the density and just concentrates on separating various regions of the space.

Classification as a Mathematical Technique:

1. The linear programming formulation obtains an approximate separating plane that minimizes a weighted sum of the distances of misclassified points to the approximate separating plane. Minimization of such a weighted sum of the distances of misclassified points by a linear program is merely a surrogate for minimizing the number of misclassified points by the separating plane.
2. A precise mathematical programming formulation of the nonconvex problem of minimizing the number of such misclassified points [8]. This is done by first proposing a simple linear complementarity formulation of the step function (Lemma 2.1) and then using this result to formulate the misclassification minimization problem as a linear program with equilibrium (linear complementarities) constraints (LPEC).
3. A linear program with equilibrium constraints (LPEC) is a linear program with a single complemetarity constraint (an orthogonality condition between two linear functions). LPECs arise when the constraints involve another linear programming problem. LPECs can model machine learning problems [8, 9], while more general mathematical programs with equilibrium constraints (MPECs) [10] can model more general problems such as economic and traffic equilibrium problems.

2. Regression

The regression problem differs from the classification problem in this the function g has a continuous output in contrast to a discrete output for the classification

Regression as a Mathematical Technique:

*a. Linear Regression Analysis:*Linear regression is an approach to modeling the relationship between a scalar dependent variable *y* and one or more explanatory variables denoted *X*. The case of one explanatory variable is called *simple linear regression*. For more than one explanatory variable, it is called *multiple linear regression*.

b. Simple Linear Regression Analysis: Simple linear regression is the least squares estimator of a linear regression model with a single explanatory variable. In other words, simple linear regression fits a straight line through the set of *n* points in such a way that makes the sum of squared *residuals* of the model as small as possible.

*c. Nonlinear Regression Analysis:*Nonlinear regression is a form of regression analysis in which observational data are modeled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables. The data are fitted by a method of successive approximations.

3. Clustering

Given a finite sampling of points (or database) from a space X, the target of clustering or segmentation is to group the data into sets of "*like*" points. The goal being to obtain clusters which provide a high-level characterization of points belonging to an individual cluster. For instance, a cluster of similar objects may turn out to share a common cause or elements of a given cluster may relate to some important goal [11]. In the clustering problem, we are attempting to definea

International Journal of Computing Algorithm
Volume: 03 Issue: 03 December 2014 Pages: 260-263
ISSN: 2278-2397

"*useful*" classification function over the set. Clustering algorithms typically employ a two-stage search: An outer loop over possible cluster numbers and an inner loop to fit the best possible clustering for a given number of clusters. Given the number k of clusters, clustering methods can be divided into three classes:

- *Metric-distance based methods:* a distance measure is defined and the objective becomes finding the best k-way partition such that cases in each block of the partition are closer to each other (or centroid) than to cases in other clusters.
- *Model-based methods:* a model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each cluster.
- *Partition-based methods:* basically enumerate various partitions and then score them by some criterion. The above two techniques can be viewed as special cases of this class. Many techniques in the AI literature fall under this category and utilize ad hoc scoring functions.

Clustering as a Mathematical Technique:

1. *Connectivity based clustering:* Connectivity based clustering, also known as hierarchical clustering, is based on the core idea of objects being more related to nearby objects than to objects farther away. These algorithms connect "objects" to form "clusters" based on their distance.
2. *Centric-based clustering:*In centric-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to k, *k*-means clustering gives a formal definition as an optimization problem: find the k cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized.
3. *Distribution based clustering:*The clustering model most closely related to statistics is based on distribution models. Clusters can then easily be defined as objects belonging most likely to the same distribution. A nice property of this approach is that this closely resembles the way artificial data sets are generated: by sampling random objects from a distribution.
4. *Density based clustering:*In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points. The most popular density based clustering method is DBSCAN.

4. Dependency Modeling

Insight into data is often gained by deriving some causal structure within the data. Models of causality can be probabilistic as in deriving some statement about the probability distribution governing the data or they can be deterministic as in deriving functional dependencies between fields in the data [12]. Density estimation methods in general fall under this category, so do methods for explicit causal modeling.

Dependency Modeling as Mathematical Technique:

- *Fault Tree Analysis* has become an established tool used to assess the likelihood of failure of industrial systems. It is particularly well utilized for the assessment of safety systems whose failure can cause excessive financial loss. The popularity of the method is due to the system failure is represented in a logic tree diagram and the terminal branch represents the component failure. This form of diagram represents a mathematical logic equation, which provides the documented means of representing the fault propagation.

5. Data Summarization

Sometimes the goal of a data mining method is to simply extract compact patterns that describe subsets of the data. There are two classes of methods which represent taking horizontal (cases) or vertical (fields) slices of the data. In the former, one would like to produce summaries of subsets: e.g. producing sufficient statistics, or logical conditions that hold for subsets. In the latter case, one would like to predict relations between fields. This class of methods is distinguished from the other data mining methods discussed in that rather than predicting a specified field (e.g. classification) or grouping cases together (e.g. clustering) the goal is to find relations between fields. One common method is by *association rules* [13]. Associations are rules that state that certain combinations of values occur with other combinations of values with a certain frequency and certainty. A common application of this is market basket analysis were one would like to summarize which products are bought with what other products. While there are exponentially many rules, due to data sparseness only few such rules satisfy given support and confidence thresholds. Scalable algorithms find all such rules in linear time (for reasonable threshold settings). While these rules should not be viewed as statements about causal effects in the data, they are useful for modeling purposes if viewed as frequent marginal in a discrete (e.g. 908multinomial) probability distribution. Of course to do proper inference one needs to know the frequent, infrequent, and all probabilities in between. However, approximate inference can sometimes be useful.

6. Change and Deviation Detection

These methods account for sequence information, be it time-series or some other ordering (e.g. protein sequencing in genome mapping). The distinguishing feature of this class of methods is that ordering of observations is important and must be accounted for. Scalable methods for finding frequent sequences in databases, while in the worst-case exponential in complexity, do appear to execute efficiently given sparseness in real-world transactional databases [14].

III. CONCLUSION

Data mining and KDD holds the promise of an enabling technology that could reveal the knowledge hidden in huge databases. Perhaps the most exciting aspect is the possibility of the evolution of new methods properly mixing Mathematics, Statistics, Databases, Optimization, Automated Data Analysis and Reduction, and other related areas. This mix would produce new algorithms and methodologies, tuned to working on large databases and scalable both in data set size and in parallelism. This paper provides an overview of this area defined in basic terms with some of the basic mathematical techniques which can be applied in data mining concept. The

International Journal of Computing Algorithm
Volume: 03 Issue: 03 December 2014 Pages: 260-263
ISSN: 2278-2397

suitable blend of these techniques will greatly useful in exploring the massive, ever growing and ever changing datasets.

REFERENCES

[1] Edward E. Leamer. Specification searches: ad hoc inference with nonexperimental data. Wiley, New York, 1978.
[2] G. Piatetsky-Shapiro and W. Frawley, editors.Knowledge Discovery in Databases.MIT Press, Cam-bridge, MA, 1991.
[3] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: An overview. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, Advances in knowledgeDiscovery and Data Mining, pages 1 – 36. MIT Press, Cambridge, MA, 1996.
[4] R.O. Duda and P.E. Hart.Pattern Classification and Scene Analysis.John Wiley and Sons, New York, 1973.
[5] D. Heckerman. Bayesian networks for data mining. Data Mining and Knowledge Discovery, 1(1), 1997.
[6] J. Hertz, A. Krogh, and R. G. Palmer.Introduction to the Theory of Neural Computation.Addison-Wesley, Redwood City, California, 1991.
[7] O. L. Mangasarian. Multi-surface method of pattern separation.IEEE Transactions on InformationTheory, IT-14:801–807, 1968.
[8] O. L. Mangasarian. Misclassification minimization.Journal of Global Optimization, 5:309–323, 1994.
[9] O. L. Mangasarian. Mathematical programming in machine learning.In G. Di Pillo and F. Giannessi, editors, Nonlinear Optimization and Applications, pages 283–295, New York, 1996.Plenum Publishing.
[10] Z.-Q. Luo, J.-S.Pang, and D. Ralph. Mathematical Programs with Equilibrium Constraints.Cambridge University Press, Cambridge, England, 1996.
[11] J. W. Shavlik and T. G. Dietterich (editors).Readings in Machine Learning.Morgan Kaufman, San Mateo, California, 1990.
[12] G. Piatetsky-Shapiro and W. Frawley, editors.Knowledge Discovery in Databases.MIT Press, Cam-bridge, MA, 1991.
[13] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and I.C. Verkamo.Fast discovery of association rules. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, Advances in knowledgeDiscovery and Data Mining, pages 307 – 328. MIT Press, Cambridge, MA, 1996.
[14] H. Mannila, H. Toivonen, and A.I. Verkamo.Discovery of frequent episodes in event sequence.DataMining and Knowledge Discovery, 1(3), 1997.
[15] *en.wikipedia.org/wiki/Mathematical_model*