International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 5, Issue 2, March - April 2016 **ISSN 2278-6856**

EXPLORING DATA MINING TECHNIQUES AND ITS APPLICATIONS

Dr.T.Hemalatha[#1], Dr.G.Rashita Banu[#2], Dr.Murtaza Ali[#3]

[#1].Assisstant.Professor,VelsUniversity,Chennai
[#2]Assistant Professor,Department of HIM&T,JazanUniversity,Jasan
[#3]HOD, Department of HIM&T JazanUniversity,Jasan

Abstract: ***Data mining is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large databases. The patterns must be actionable so that they may be used in an enterprise's decision making. There are various data mining techniques which can be implemented to achieve the desired solution. This paper focuses on various data mining techniques which will provide the necessary effectiveness in the task undertaken.***

Keywords:- Data mining, patterns, decision making, Techniques.

I. INTRODUCTION

Data mining is a powerful new technology with great potential to help companies focus on the most important information in the data they have collected about the behaviour of their customers and potential customers. It discovers information within the data that queries and reports can't effectively reveal. Data mining, or knowledge discovery, is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict behaviours and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.Data mining derives its name from the similarities between searching for valuable information in a large database and mining a mountain for a vein of valuable ore. It allows 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 software analyzes relationships and patterns in stored transaction data based on open-ended user queries.

II. DATA MINING TECHNIQUES

There are several major ***data mining techniques***have been developed and used in data mining projects recently including association, classification, clustering, prediction and sequential patterns. We will briefly examine those data mining techniques with example to have a good overview of them.

ASSOCIATION

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. For example, the association technique is used in *market basket analysis* to identify what products that customers frequently purchase together. Based on this data businesses can have corresponding marketing campaign to sell more products to make more profit. Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An example of an association rule would be "If a customer buys a dozen eggs, he is 80% likely to also purchase milk."

An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent.

Association rules are created by analyzing data for frequent if/then patterns and using the criteria *support* and *confidence* to identify the most important relationships.*Support* is an indication of how frequently the items appear in the database. *Confidence* indicates the number of times the if/then statements have been found to be true.

In data mining, association rules are useful for analyzing and predicting customer behavior. They play an important part in shopping basket data analysis, product clustering, catalog design and store layout.

Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time.

CLASSIFICATION

Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we make the software that can learn how to

classify the data items into groups. For example, we can apply classification in application that "given all past records of employees who left the company, predict which current employees are probably to leave in the future." In this case, we divide the employee's records into two groups that are "leave" and "stay". And then we can ask our data mining software to classify the employees into each group. Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.

A classification task begins with a data set in which the class assignments are known. For example, a classification model that predicts credit risk could be developed based on observed data for many loan applicants over a period of time. In addition to the historical credit rating, the data might track employment history, home ownership or rental, years of residence, number and type of investments, and so on. Credit rating would be the target, the other attributes would be the predictors, and the data for each customer would constitute a case.

Classifications are discrete and do not imply order. Continuous, floating-point values would indicate a numerical, rather than a categorical, target. A predictive model with a numerical target uses a regression algorithm, not a classification algorithm.

The simplest type of classification problem is binary classification. In binary classification, the target attribute has only two possible values: for example, high credit rating or low credit rating. Multiclass targets have more than two values: for example, low, medium, high, or unknown credit rating.

Testing a Classification Model

A classification model is tested by applying it to test data with known target values and comparing the predicted values with the known values.

The test data must be compatible with the data used to build the model and must be prepared in the same way that the build data was prepared. Typically the build data and test data come from the same historical data set. A percentage of the records is used to build the model; the remaining records are used to test the model.

Test metrics are used to assess how accurately the model predicts the known values. If the model performs well and meets the business requirements, it can then be applied to new data to predict the future.

Accuracy

Accuracy refers to the percentage of correct predictions made by the model when compared with the actual classifications in the test data.

CLUSTERING

Clustering is a data mining technique that makes meaningful or useful cluster of objects that have similar characteristic using automatic technique. Different from classification, clustering technique also defines the classes and put objects in them, while in classification objects are assigned into predefined classes. To make the concept clearer, we can take library as an example. In a library, books have a wide range of topics available. The challenge is how to keep those books in a way that readers can take several books in a specific topic without hassle. By using clustering technique, we can keep books that have some kind of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in a topic, he or she would only go to that shelf instead of looking the whole in the whole library. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statisticaldata analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

GENERAL TYPES OF CLUSTERS

A. Well- separated clusters

A cluster is a set of points so that any point in a cluster is nearest (or more similar) to every other point in the cluster as compared to any other point that is not in the cluster.

B. Center -based clusters

A cluster is a set of objects such that an object in a cluster is nearest (more similar) to the "center" of a cluster, than to the center of any other cluster. The center of a cluster is often a centroid.

C. Contiguous clusters

A cluster is a set of points so that a point in a cluster is nearest (or more similar) to one or more other points in the cluster as compared to any point that is not is the cluster

D. Density- based clusters

A cluster is a dense region of points, which is separated by according to the low density regions, from other regions that is of high density.

E. Shared Property

Finds clusters that share some common property or represent a particular concept.

PREDICTION

The prediction as it name implied is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables. For instance, prediction analysis technique can be used in sale to predict profit for the future if we consider sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction.

SEQUENTIAL PATTERNS

Sequential patterns analysis in one of data mining technique that seeks to discover similar patterns in data transaction over a business period. The uncover patterns are used for further business analysis to recognize relationships among data.

III.DATA MINING APPLICATIONS

Data mining is used for variety of applications.They are listed here

1. Relationship Marketing

Data Mining can help in analyzing customer profiles, discovering sales triggers, and in identifying critical issues that determine client loyalty and in customer retention.

2. Website design and Promotion

Web mining is used to discover the user navigation in a website and the results can help in enhancing the site design and making it more citable on the web.

3. Forensics

Data mining finds its varied application in the field of forensics. In this field it is used to detect unusual cases like frauds and defaulters in credit cards, applying biometric techniques in recognizing criminals etc.

4. Customer profiling

Profiling can help an enterprise identify its most valuable customers so that the enterprise can differentiate the needs and values of the customer.

5. Prediction and description

Data mining techniques are used for sales forecasting and analysis. It also involves selecting the attributes of the objects available in a database to predict other variables of interest.

IV.CONCLUSION

This paper analyses the various data mining techniques which can be incorporated in the research work performed under various domains to derive the optimum results.These techniques minimises the weaknesses and enhances the strength of the desired research solution. The applications of data mining is not constrained to the listed fields but iits harnessing power is unlimited.

REFERENCES

1.G.K.Gupta,Introduction to Data Mining with Case Studies,Secondedition,PHI Learning Private Limited,2011.
2.http://www.kdnuggets.com
3. Zhu, Xingquan; Davidson, Ian (2007). Knowledge Discovery and Data Mining: Challenges and Realities. New York, NY: Hershey. pp. 31–48. ISBN 978-1-59904-252-7.
4. Chen, Yudong; Zhang, Yi; Hu, Jianming; Li, Xiang "Traffic Data Analysis Using Kernel PCA and Self-Organizing Map". IEEE Intelligent Vehicles Symposium. 2006.
5. Zhu, Xingquan; Davidson, Ian. Knowledge Discovery and Data Mining: Challenges and Realities. New York, NY: Hershey. p. 18.ISBN 978-1-59904-252-7,2007.
6. Herschel, Gareth; Magic Quadrant for Customer Data-Mining Applications, Gartner Inc., 1 July 2008
7.http://www.thearling.com/text/dmtechn ques/dmtechniques.htm.
8.http://www.obgyn.cam.ac.uk/camonly/statsbook/stdatmi n.html