Learning Data Mining Techniques

Aashaykumar Dubey
Pursuing B.E.,
Student, Dwarkadas J.Sanghvi College of Engineering

Saurabh Kamath
Pursuing B.E.,
Student, Dwarkadas J.Sanghvi College of Engineering

Dhruv Kanakia
Pursuing B.E.,
Student, Dwarkadas J.Sanghvi College of Engineering

ABSTRACT
In the internet world data is on the rise. The data which emerges from the internet is huge and highly unstructured. This data can be arranged in sophisticated manner by applying various data mining techniques. This paper focuses on a number of data and text mining techniques. These techniques are applied in highly complex business problems to extract chunks of information from data which at first sight seem to have no meaning. In an uncertain and highly competitive business environment, efficiency and speed are not the only deciding factor for a business to excel. Apart from business in particular, data mining is applied in fields including weather forecasting, health and other fields where managing data is a top priority.

General Terms
ANN, CDF

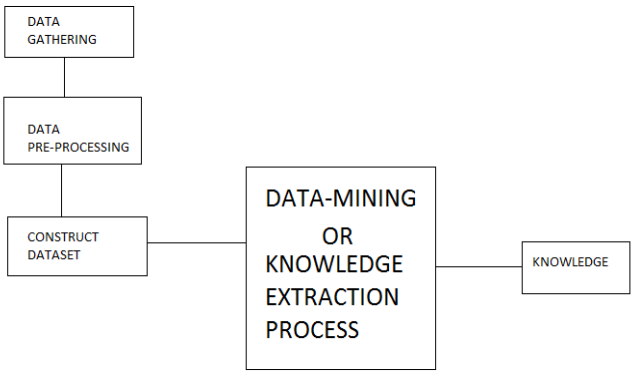
Keywords
Association, Classification, Neural networks, Decision Trees.

1. INTRODUCTION
Extracting significant information from huge collection of data which can be in various forms is the process called data mining. These forms include binary text, numbers, and texts in the form of consumer data, his /her interests or data in any other particular form. A data analysts' job is to find meaning in the data which doesn't make sense to any other person.

The enormous improvements in information systems which stores data in the form of millions of bytes has made it possible to manage data in a rich and feasible manner. Exploiting the data accumulated in this system to extract useful and actionable information, is the ultimate goal of the generic activity termed as data mining.

It can be defined as a predetermined set of rules is run against a collection of unordered and unorganized data so as to extract meaningful patterns and structures using various SEO tools or semi-automated procedures

It is a growing subfield of computer science which involves computational processing of large data sets' pattern discovery.



The advancements made in database and the invention of high storage devices have resulted in creation of large amounts of data which makes it imperative to conduct text and data mining to extract useful data. Data mining, briefly, is a process of extraction of useful information and patterns from huge data. Knowledge extraction or knowledge mining are the other names for data mining.

Data mining is also stated as required procedure where artificial intelligence methods are applied in order to extract the data patterns. Data mining consists of four major points:

- Manage exhaustive database systems.
- Provide complete access to business analysts to create data models for market analysis.
- Analyze the data by application software.
- Convert unstructured data into graphs, tables and various other structures.

2. DESCRIPTIVE TECHNIQUES
2.1 Association
Association is a data mining technique in which the relation observed in the unorganized data is used to extract useful data. It is a most common technique in online stores and is the most simplest of its kind. This is nothing but association based on relation which helps in making new marketing strategies.

For example, the association technique is used in shopping cart analysis to identify what products that customers frequently purchase together. Based on this data online marketplaces can launch corresponding marketing campaigns to boost their market shares. It is also used extensively in online advertisements. If a person visits a certain category of websites, particular cookies data can be taken as an input from the user's private computer to launch adaptive advertisements.

Market Basket Analysis

PRODUCT PURCHASED	COMPLEMENTARY PRODUCT
A	B
B	C
A & B	C

98% OF THE PEOPLE WHO BOUGHT PRODUCT A & B ALSO BOUGHT PRODUCT C.

Different association rules are based on Quantitative analysis and Boolean algebra. Association rules are used to fulfil both support and confidence criteria of the user. There can also be other parameters such as customer age, sex, preferences etc. All these minimum threshold levels must be met simultaneously, this is an important aspect of Association.

Based on the type of analysis required a single-dimensional or multidimensional approach can be carried out. In a single-dimensional approach, only a single parameter such as the user's age is used to determine the tactics that should be used to increase profits on the online platform. In a similar manner a number of parameters can also be used.

Various Algorithms are available to implement Association technique. Apriori algorithm is widely used and is the preliminary step towards extracting data using Association technique. It uses a breadth-first search strategy (direct relation) to count the support of user sets and uses a custom function which makes full use of the bottom-up property of data. For example, the rule found in the sales data of an online store like Amazon would indicate that if a customer buys a mobile phone, they are likely to also buy phone covers and screen guards. Providing offers on such products in combo-packs to gain high profits.

2.2 Clustering

The process of organizing collection of data having some kind of similarity from a data set into member groups is called clustering. Clustering is the process of organizing objects in a data set into groups which are similar. This is very much similar to association but it does not provide a further step which an associative model is capable of. Clustering involves only the arrangement of data according to a descriptive model.

A cluster is therefore a collection of objects which are similar and are dissimilar to other cluster objects.

A library can be taken as an example. In a library, books available offer a wide range. The challenge is to keep the books in such a way that readers can make use of the books in their specific manner. By using the clustering technique, the books can be categorized in a cluster based on their labels. If readers want to read books pertaining to a certain topic, he or she would only look in a particular shelf and not in the library. This technique is now used particularly on an online marketplace where a variety of consumer goods are available but they are grouped together so that the consumer can browse through exactly the goods he or she needs and not spending time on browsing unwanted items.

3. PREDICTIVE TECHNIQUES

This technique runs an exhaustive search on the entire data set in order to find the similarities between independent and dependent objects.

The predictive analysis technique is not as easy as it may seem. Predictive analysis is used in real world problems such as stock prediction, equity markets, and both fundamental and technical mutual fund portfolio analysis and risk management. The problems in real world are not entirely dependent on just a single parameter, but are dependent on complex interactions of various variables. Therefore, more complex techniques may be necessary to forecast future stock prices.

Prediction analysis technique can be used in sale to estimate the profit considering that sale is an independent variable and profit may be a dependent variable. Then based on the sale and profit data, fitted regression curve that is used for profit estimation may be drawn.

3.1 Classification

Classification is a method which forms the basic stepping stone for various recursive algorithms implemented for data mining. The design goal of a classification algorithm is to split various data-sets or database which have huge data into several groups so that they can be analyzed using a single exhaustive model.

Classification method makes use of statistical techniques such as decision trees, linear programming, neural network and statistic-oriented approach.

In classification, a software that can learn how to classify the data items into groups can be used. For example, classification in application that predicts which employees will probably leave in the future based on the past employee record. In this case, the records are divided into two groups that are based on employees which are staying and employees which are leaving. The data mining software helps in classifying the employees into different groups.Similarly, all the customers can be classified based on the amount of purchases within a particular span of time. This allows the advertisement campaigns to focus on more active customers.

Classification Techniques are of three types:

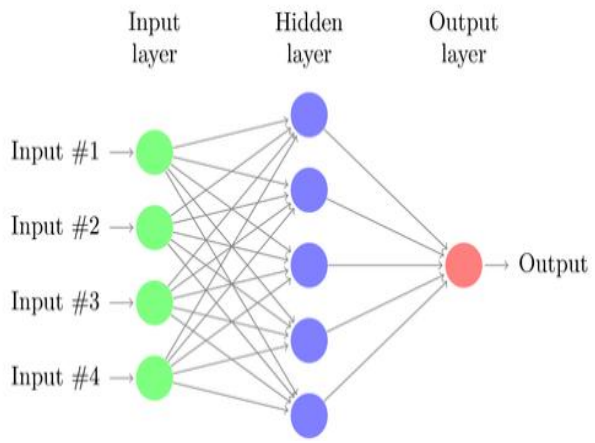
- Neural Networks
- Decision trees
- Regression

3.1.1 Neural Networks

In data mining a statistical model known as Artificial Neural Networks or simply neural network is used. It is abbreviated as ANN or NN

It consists of an interconnected group of artificial neurons and processes information using an approach to computation based on the connections. In most cases an ANN is an adaptive system that changes its form based on information that flows through the network during the learning phase.

Neural Network or an artificial neural network is a biological system that detects patterns and makes predictions. In recent years the discovery in neural network are in their application to real time problems like customer response, fraud detection etc.



Data mining techniques such as neural networks are able to model the relationships existing in data collections and can be used for increasing business intelligence across a variety of applications. This predictive modelling technique creates very complex models that are really difficult to understand, even for experts. Neural Networks are used in a variety of applications.

It is shown in figure, Artificial neural network have become a great tool in tasks like pattern recognition, decision problem or predication applications. This processing method came into action a few years back. ANN is an adaptive, nonlinear system that learns to perform a function from data is normally training phase where system parameter is change during operations. The parameters are fixed. Flexibility in a data model can be increased if the problem is poorly framed or not easy to understand then using ANN model it can be made accurate.

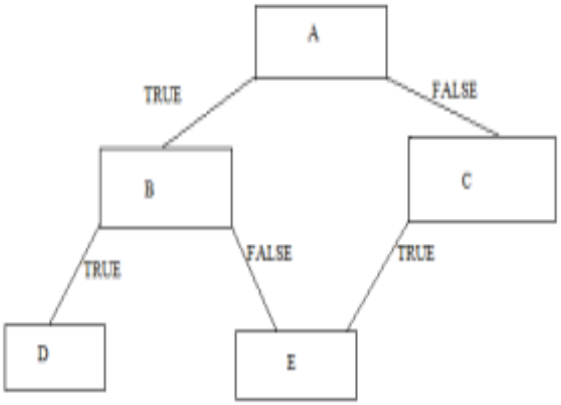
Maintaining databases using neural networks is easy. It does not require conventional storage centers, so the space and cost reduces. Neural networks can handle highly non-linear mappings and the mapping is quick and stable. This provides flexibility and large storage space. The input to the neural network on an online platform is the various parameters of the customer. Statistical models can be implemented in the hidden networks. These hidden networks are implemented in an isolated atmosphere. This is usually a firewall protected high end CPU. The output is observed at the output network. These output networks are analyzed by a data miner. Patterns can be observed on the output nodes. Usually health related information in the field of medical science is implemented using neural networks. In the recent past this technique has found applications in various other fields.

3.1.2 Decision Trees

A decision tree is a structure that includes a root node, branches, and leaf nodes. Each node depicts a test on an attribute, each branch denotes the outcome of a test under consideration, and each leaf node holds a label for a particular class. The topmost node in the tree is the root node.

A decision tree is a predictive model used for classification. Decision trees partition the input space into cells where every cell belongs to a class. The partitioning is represented as a sequence of tests such as Y=X(x1,x2,x3,.....,xn), where Y is the outcome of a particular class of test and x1,x2....xn are the inputs to the function 'X' which can be a predictive function. Predictive functions can be cumulative distribution functions (C.D.F) or probabilistic functions. A tree is traced from the root node to the leaf node by splitting the original set into subsets on the basis of attributes of a particular class. This process is repeated in the same way as the recursive Greedy Algorithm is implemented. The recursion stops the node/variable has the same value as the target node/variable. Greedy's Algorithm is also known by 'the top-down induction of Decision trees'.

Decision trees can be viewed from the business perspective as creating a segmentation of the original data set into sub-classes on the basis of their categories. Thus marketing managers use segmentation of customers, products and sales data sheets for predictive analysis.



Logical operations are used to learn a decision tree. Since probability includes a number of operators such as Logical AND, OR, EXOR, NOR etc. out of which AND, OR operators find applications in Data Mining. To join two sub-classes of a decision tree 'AND' operation is implemented. In the case of disjunction of two sub-classes, 'OR' operation is executed.

3.1.3 Regression

Regression is a data mining (machine learning) technique used to fit an equation to a dataset. The simplest form of regression uses the formula of a straight line. A straight line is given by the equation y = mx + c and determines the approximate values for m and c to calculate the value of y based on a particular value of x. Advanced techniques, such as multiple regression, uses more than one input variable and allows for the fitting of more complex models, such as a quadratic equation. Usually a straight line model is used to find the residuals or deviations in the data set. These irregularities are present because the data doesn't match exactly to the data present in the data set. In multivariate linear regression, the regression parameters referred as coefficients. While building a multivariate linear regression model, the algorithm computes a coefficient for each of the indicators of deviation used by the model. The coefficient is a measure of the impact of the movement of value x on the target y. Various methods are used to determine the values of the variables of the equation which satisfy the regression curve. This is known as "Regression Statistics". The relationship between x and y cannot be approximated with a straight line. In this case, a nonlinear regression technique may be used.

Nonlinear regression models define y as the function of x using an equation that is more complicated than the linear regression equation

4. DATA MINING APPROACH

The purpose of a data mining effort is to create a descriptive model or a predictive model.

A descriptive model enlists the important properties in the data. It is essentially a written pattern in order to describe the changes or deviations present in the data. Typically, a descriptive model is most commonly found in all mining techniques. It is a passive method which gives details about the patterns in the data but it does not provide any kind of analysis in any way. In other words this technique leaves the interpretation of the patterns to the data miner. The aim of a predictive model is to give the user the liberty to guess a specific variable.

The predictive model creates a data set which approximately defines the data available. The patterns are then compared with this model to find the deviation and is then analyzed or coded in the deviated form.

5. CONCLUSION

This paper briefly reviews the data mining techniques which are in use. This review paper provides help to an individual to focus on the various techniques developed for data mining.

The definition restricts its importance to the area in the field of computer science but it will be broadened in the future. It is also evident from the above study that designing a data mining system is not an easy task. The most important challenge data mining faces is detecting understandable patterns and to make the methods user friendly. Perhaps designing and developing a system which can work for any domain with great accuracy is extremely complicated and will have its limitations. The major goal is to achieve a user friendly transparent system with reduced parameters.

Although the future cannot be seen, many new challenges will be present which cannot be predicted at this point of time.

6. ACKNOWLEDGMENTS

We thank our teachers and our institute for encouraging us to work and develop ideas on topics as such.

7. REFERENCES

[1] Larose, D. T., “Discovering Knowledge in Data: An Introduction to Data Mining”, ISBN 0-471-66657-2, ohn Wiley & Sons, Inc., 2005.

[2] Tan Pang-Ning, Steinbach, M., Vipin Kumar. “Introduction to Data Mining”, *Pearson Education, New Delhi, ISBN: 978-81-317-1472-0, 3rd Edition, 2009.* Bernstein, A. and Provost, F., “An Intelligent Assistant for the Knowledge Discovery Process”,

[3] M. Craven and J. Shavlik, “Learning rules using ANN “, Proceeding of 10th International Conference on Machine Learning, pp.-73-80, July 1993.

[4] Lior Rokach and Oded Maimon, “Data Mining with Decision Trees: Theory and Applications (Series in Machine Perception and Artificial Intelligence)”, I SBN: 981-2771-719, World Scientific Publishing Company, 2008.

[5] "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients." Expert Systems with Applications 36(2), 2473–2480.

[6] Venkatadri.M and Lokanatha C. Reddy ,“A comparative study on decision tree classification algorithm in data mining” , International Journal Of Computer Applications In Engineering ,Technology And Sciences, Sept 2010.