DATA MINING: PROBLEMS FACED IN RESEARCH PAPER

[1]P. Shiamilee, [2]S. Aron Raj, [3]D. Richard
[1]Student, [2]Student, [3]Assistant Professor
[1]Information Technology,
[1]St.Joseph's College(Autonomous), Trichy, India

Abstract : ***In this paper, the concept of data mining was summarized and its significance towards its methodologies and researches issues faced my researchers was illustrated. The data mining based on Neural Network and Genetic Algorithm is researched in detail and the key technology and ways to achieve the data mining on Genetic Algorithm are also surveyed. It took an initiative to identify some challenging problems in data mining research, by referencing some of the most active researcher's paper in data mining and machine learning for their opinions on what are considered important and worthy topics for future research in data mining. This short article serves to summarise the most challenging problems of the data mining. The order of the listing does not reflect their level of importance.***

IndexTerms - Data Mining, Machine learning, knowledge discovery, Genetic Algorithm, methodologies of data mining.

I. INTRODUCTION

Data mining refers to extracting or mining the knowledge from large amount of data. The term data mining is appropriately named as 'Knowledge mining from data' or "Knowledge mining". Data collection and storage technology has made it possible for organizations to accumulate huge amounts of data at lower cost. Exploiting this stored data, in order to extract useful and actionable information, is the overall goal of the generic activity termed as data mining. The following definition is given:
Data mining is the process of exploration and analysis, by Automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules.

In [1], the following definition is given: Data mining is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules. Data mining is an interdisciplinary subfield of computer science which involves computational process of large data sets' patterns discovery. The goal of this advanced analysis process is to extract information from a data set and transform it into an understandable structure for further use. The methods used are at the juncture of artificial intelligence, machine learning, statistics, database systems and business intelligence. Data Mining is about solving problems by analysing data already present in databases [2].
Data mining is also stated as essential process where intelligent methods are applied in order to extract the data patterns.

Data mining consists of five major elements:
Extract, transform, and load transaction data onto the data warehouse system.
Store and manage the data in a multidimensional database system.
Provide data access to business analysts and Information technology professionals.
Analyse the data by application software.
Present the data in a useful format, such as a graph or table.
Data mining functionalities are used to specify the kind of Patterns to be found in data mining tasks. Data mining tasks can be classified in two categories-descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in database. Predictive mining tasks perform inference on the current data in order to make predictions. The purpose of a data mining effort is normally either to create a descriptive model or a predictive model.

A descriptive model presents, in concise form, the main characteristics of the data set. It is essentially a summary of the data points, making it possible to study important aspects of the data set. Typically, a descriptive model is found through undirected data mining; i.e. a bottom-up approach where the data "speaks for itself". Undirected data mining finds patterns in the data set but leaves the interpretation of the patterns to the data miner. The purpose of a predictive model is to allow the data miner to predict an unknown (often future) value of a specific variable, the target variable. If the target value is one of a predefined number of discrete (class) labels, the data mining task is called classification. If the target variable is a real number, the task is regression. The predictive model is thus created from given known values of variables, possibly including previous values of the target variable. The training data consists of pairs of measurements, each consisting of an input vector x (i) with a corresponding Target value y(I). The predictive model is an estimation of the function y=f(x; q) able to predict a value y, given an input vector of measured values x and a set of estimated parameters q for the model f. The process of finding the best q values is the core of the data mining technique [3].

At the core of the data mining process is the use of a data mining technique. Some data mining techniques directly obtain the information by performing a descriptive partitioning of the data. More often, however, data mining techniques utilize stored data in order to build predictive models. From a general perspective, there is strong agreement among both researchers and executives about the criteria that all data mining techniques must meet. Most importantly, the techniques should have high performance. This criterion is, for predictive modelling, understood to mean that the technique should produce models that will generalize well, i.e. models having high accuracy when performing predictions based on novel data. Classification and prediction are two forms of data analysis that can be used to extract models describing the important data classes or to predict the future data trends. Such analysis can help to provide us with a better understanding of the data at large. The classification predicts categorical (discrete, unordered) labels, prediction model, and continuous valued Function.

II. RESEARCH PROBLEMS IN DATA MINING

2.1 Scaling Up for High Dimensional Data and High Speed Data Streams:

One challenge is how to design classifiers to handle ultra-high dimensional classification problems. There is a strong need now to build useful classifiers with hundreds of millions or billions of features, for applications such as text mining and drug safety analysis. Such problems often begin with tens of thousands of features and also with interactions between the features, so the number of implied features gets huge quickly. One important problem is mining data streams in extremely large databases (e.g. 100 TB). Satellite and computer network data can easily be of this scale. However, today's data mining technology is still too slow to handle data of this scale. In addition, data mining should be a continuous, on line process, rather than an occasional one-shot process. Organizations that can do this will have a decisive advantage over ones that do not. Data streams present a new challenge for data mining researchers. One particular instance is from high speed network traffic where one hopes to mine information for various purposes, including identifying anomalous events possibly indicating attacks of one kind or another. A technical problem is how to compute models over streaming data, which accommodate changing environments from which the data are drawn. This is the problem of "concept drift" or "environment drift." This problem is particularly hard in the context of large streaming data. How may one compute models that are accurate and useful very efficiently? For example, one cannot presume to have a great deal of computing power and resources to store a lot of data, or to pass over the data multiple times. Hence, incremental mining and effective model updating to maintain accurate modelling of the current stream are both very hard problems. Data streams can also come from sensor networks and RFID applications. In the future, RFIDs will be a huge area, and analysis of this data is crucial to its success.

2.2 Mining Sequence Data and Time Series Data:

Sequential and time series data mining remains an important problem. Despite progress in other related fields, how to efficiently cluster, classify and predict the trends of these data is still an important open topic.

A particularly challenging problem is the noise in time series data. It is an important open issue to tackle. Many time series used for predictions are contaminated by noise, making it difficult to do accurate short-term and long-term predictions. Examples of these applications include the predictions of financial time series and seismic time series. Although signal processing techniques, such as wavelet analysis and filtering, can be applied to remove the noise, they often introduce lags in the filtered data. Such lags reduce the accuracy of predictions because the predictor must overcome the lags before it can predict into the future. Existing data mining methods also have difficulty in handling noisy data and learning meaningful information from the data.
Some of the key issues that need to be addressed in the design of a practical data miner for noisy time series include:

Information/search agents to get information: Use of wrong, too many, or too little search criteria; possibly inconsistent information from many sources; semantic analysis of (meta-) information; assimilation of information into inputs to predictor agents.
Learner/miner to modify information selection criteria: Apportioning of biases to feedback; developing rules for Search Agents to collect information; developing rules for Information Agents to assimilate information.
Predictor agents to predict trends: Incorporation of qualitative information; multi objective optimization not in closed form.

2.3 Mining Complex Knowledge from Complex Data:

One important type of complex knowledge is in the form of graphs. Recent research has touched on the topic of discovering graphs and structured patterns from large data, but clearly, more needs to be done. Another form of complexity is from data that are non-i.i.d. (independent and identically distributed). This problem can occur when mining data from multiple relations. In most domains, the objects of interest are not independent of each other, and are not of a single type. We need data mining systems that can soundly mine the rich structure of relations among objects, such as interlinked Web pages, social networks, metabolic networks in the cell, etc. Yet another important problem is how to mine non-relational data.

A great majority of most organizations' data is in text form, not databases, and in more complex data formats including Image, Multimedia, and Web data. Thus, there is a need to study data mining methods that go beyond classification and clustering. Some interesting questions include how to perform better automatic summarization of text and how to recognize the movement of objects and people from Web and Wireless data logs in order to discover useful spatial and temporal knowledge. There is now a strong need for integrating data mining and knowledge inference. It is an important future topic. In particular, one important area is to incorporate background knowledge into data mining.

The biggest gap between what data mining 600 Q. Yang & X. Wu systems can do today and what we'd like them to do is that they're unable to relate the results of mining to the real-world decisions they affect all they an do is hand the results back to the user. Doing these inferences, and thus automating the whole data mining loop, requires representing and using world knowledge within the system. One important application of the integration is to inject domain information and business knowledge into the knowledge discovery process. Related to mining complex knowledge, the topic of mining interesting knowledge remains important. In the past, several researchers have tackled this problem from different angles, but we still do not have a very good understanding of what makes discovered patterns "interesting" from the end-user perspective.

2.4 Security, Privacy, and Data Integrity:

Several researchers considered privacy protection in data mining as an important topic. That is, how to ensure the users' privacy while their data are being mined. Related to this topic is data mining for protection of security and privacy. One respondent states that if we do not solve the privacy issue, data mining will become a derogatory term to the general public. Some respondents consider the problem of knowledge integrity assessment to be important. We quote their observations: "Data mining algorithms are frequently applied to data that have been intentionally modified from their original version, in order to misinform the recipients of the data or to counter privacy and security threats. Such modifications can distort, to an unknown extent, the knowledge contained in the original data.

As a result, one of the challenges facing researchers is the development of measures not only to evaluate the knowledge integrity of a collection of data, but also of measures to evaluate the knowledge integrity of individual patterns. Additionally, the problem of knowledge integrity assessment presents several challenges." Related to the knowledge integrity assessment issue, the two most significant challenges are: (1) develop efficient algorithms for comparing the knowledge contents of the two (before and after) versions of the data, and (2) develop algorithms for estimating the impact that certain modifications of the data have on the statistical significance of individual patterns obtainable by broad classes of data mining algorithms.

The first challenge requires the development of efficient algorithms and data structures to evaluate the knowledge integrity of a collection of data. The second challenge is to develop algorithms to measure the impact that the modification of data values has on a discovered pattern's statistical significance, although it might be infeasible to develop a global measure for all data mining algorithms.

III. METHODOLOGIES OF DATA MINING

3.1 Neural Network:

Neural Network or an artificial neural network is a biological system that detects patterns and makes predictions. The greatest breakthroughs in neural network in recent years are in their application to real world problems like customer response prediction, fraud detection etc. Data mining techniques such as neural networks are able to model the relationships that exist in data collections and can therefore be used for increasing business intelligence across a variety of business applications [4]. This powerful predictive modelling technique creates very complex models that are really difficult to understand by even experts. Neural Networks are used in a variety of applications. It is shown in fig.1. Artificial neural network have become a powerful tool in tasks like pattern recognition, decision problem or predication applications. It is one of the newest signals processing technology. ANN is an adaptive, non-linear system that learns to perform a function from data and that adaptive phase is normally training phase where system parameter is change during operations. After the training is complete the parameter are fixed. If there are lots of data and problem is poorly understandable then using ANN model is accurate, the nonlinear characteristics of ANN provide it lots of flexibility to achieve input output map. Artificial Neural Networks, provide user the capabilities to select the network topology, performance parameter, learning rule and stopping criteria.

.

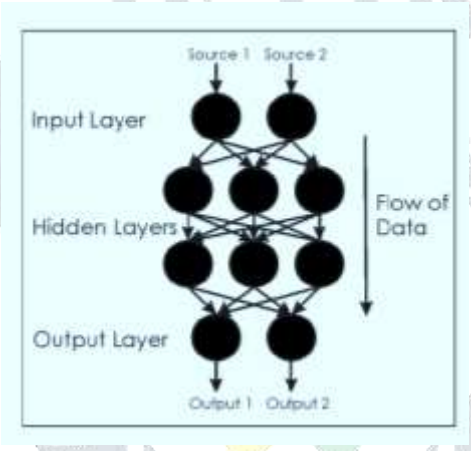

FIGURE: NEURAL NETWORK WITH HIDDEN LAYERS

3.2 Decision Trees:

A decision tree is a flow chart like structure where each node denotes a test on an attribute value, each branch represents an outcome of the test and tree leaves represent classes or class distribution. A decision tree is a predictive model most often used for classification. Decision trees partition the input space into cells where each cell belongs to one class. The partitioning is represented as a sequence of tests. Each interior node in the decision tree tests the value of some input variable, and the branches from the node are labelled with the possible results of the test. The leaf nodes represent the cells and specify the class to return if that leaf node is reached. The classification of a specific input instance is thus performed by starting at the root node and, depending on the results of the tests, following the appropriate branches until a leaf node is reached [5]. Decision tree is represented in figure below

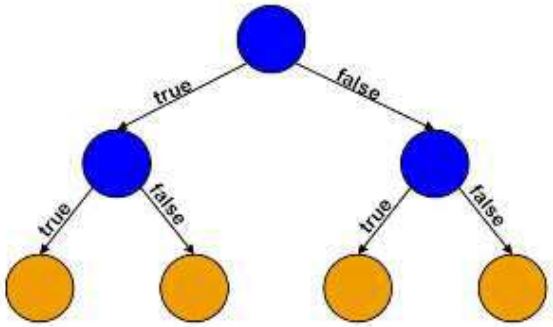

Figure: Decision Tree

Decision tree is a predictive model that can be viewed as a tree where each branch of the tree is a classification question and leaves represent the partition of the data set with their classification. The author defines a Decision Tree as a schematic tree-shaped diagram used to determine a course of action or show a statistical probability [6]. Decision trees can be viewed from the business perspective as creating a segmentation of the original data set. Thus marketing managers make use of segmentation of customers, products and sales region for predictive study. These predictive segments derived from the decision tree also come with a description of the characteristics that define the predictive segment. Because of their tree structure and skill to easily generate rules the method is a favoured technique for building understandable models.

3.3 Genetic Algorithm

Genetic Algorithm attempt to incorporate ideas of natural evaluation The general idea behind GAs is that they can build a better solution if they somehow combine the "good" parts of other solutions (schema theory), just like nature does by combining the DNA of living beings [7]. Genetic Algorithm is basically used as a problem solving strategy in order to provide with a optimal solution. They are the best way to solve the problem for which little is known. This will work well in any search space because they form a very general algorithm. The only thing to be known is what the particular situation is where the solution performs very well, and a genetic algorithm will generate a high quality solution. Genetic algorithms use the principles of selection and evolution to produce several solutions to a given problem. It is shown in fig.

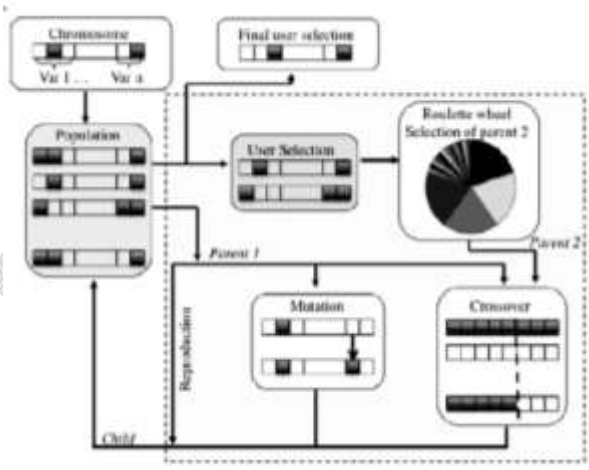

Figure : Structural view of Genetic Algorithm

Genetic algorithms (GAs) [8] are based on a biological applications; it depends on theory of evolution. When GAs are used for problem solving, the solution has three distinct stages:

The solutions of the problem are encoded into representations that support the necessary variation and selection operations; these representations, are called chromosomes, are as simple as bit strings.

A fitness function judges which solutions are the “best” life forms, that is, most appropriate for the solution of the particular problem. These individuals are favoured in survival and reproduction, thus giving rise to generation.

Crossover and mutation produce new gene individuals by recombining features of their parents. Eventually a generation of individuals will be interpreted back to the original problem domain and the fit individual represents the solution.

IV. CONCLUSION

Data mining has achieved tremendous success. Many new problems have emerged and have been solved by data mining researchers. However, there is still a lack of timely exchange of important topics in the community as a whole. This article summarizes a survey that we have conducted to rank some most important problems in data mining research. These problems are sampled from a small, important, segment of the community. The list should obviously be a function of time for this dynamic field. Finally, the source summarizes some problems below:

Scaling up for high dimensional data and high speed data streams
Mining sequence data and time series data
Mining complex knowledge from complex data
Data mining in a network setting
Distributed data mining and mining multi-agent data
Data mining process-related problems
Security, privacy and data integrity

If the conception of computer algorithms being based on the evolutionary of the organism is surprising, the extensiveness with which these methodologies are applied in so many areas is no less than astonishing. At present data mining is a new and important area of research and ANN itself is a very suitable for solving the problems of data mining because its characteristics of good robustness, self-organizing adaptive, parallel processing, distributed storage and high degree of fault tolerance. The commercial, educational and scientific applications are increasingly dependent on these methodologies.

REFERENCES

[1] Xingquan Zhu, Ian Davidson, "Knowledge Discovery and Data Mining: Challenges and Realities", ISBN 978- 1-59904-252, Hershey, New York, 2007.

[2] Joseph, Zernik, "Data Mining as a Civic Duty – Online Public Prisoners Registration Systems", International Journal on Social Media: Monitoring, Measurement, Mining, vol. - 1, no.-1, pp. 84-96, September 2010.

[3] Zhao, Kaidi and Liu, Bing, Tirpark, Thomas M. and Weimin, Xiao,"A Visual Data Mining Framework for Convenient Identification of Useful Knowledge", ICDM '05 Proceedings of the Fifth IEEE International Conference on Data Mining, vol.-1, no.-1,pp.- 530- 537,Dec 2005.

[4] R. Andrews, J. Diederich, A. B. Tickle," A survey and critique of techniques for extracting rules from trained artificial neural networks", Knowledge-Based Systems, vol.- 8,no.-6, pp.-378-389,2012.

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

[6] M. Venkatadri 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 (IJCAETS), Vol.- 2 ,no.- 2 , pp. 24- 29 , Sept 2010.

[7] Ankita Agarwal,"Secret Key Encryption algorithm using genetic algorithm", vol.-2, no.-4, ISSN: 2277 128X, IJARCSSE, pp. 57-61, April 2012.

[8] Li Lin, Longbing Cao, Jiaqi Wang, Chengqi Zhang, "The Applications of Genetic Algorithms in Stock Market Data Mining Optimisation", Proceedings of Fifth International Conference on Data Mining, Text Mining and their Business Applications, pp. 593-604, Sept 2013.