DATA MINING TECHNIQUES: A SURVEY PAPER

Nikita Jain[1], Vishal Srivastava[2]

[1]M. Tech. Scholar, [2]Associate Professor, Arya College of Engineering and IT, Rajasthan, India,
nikitagoodjain@gmail.com, vishal500371@yahoo.co.in

Abstract

In this paper, the concept of data mining was summarized and its significance towards its methodologies 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 Neural Network and Genetic Algorithm are also surveyed. This paper also conducts a formal review of the area of rule extraction from ANN and GA.

Keywords**:** *Data Mining, Neural Network, Genetic Algorithm, Rule Extraction.*

---***---

1. 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 semiautomatic 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 analyzing 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.
- Analyze 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.

2. METHODOLOGIES OF DATA MINING

2.1Neural 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 non linear 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.

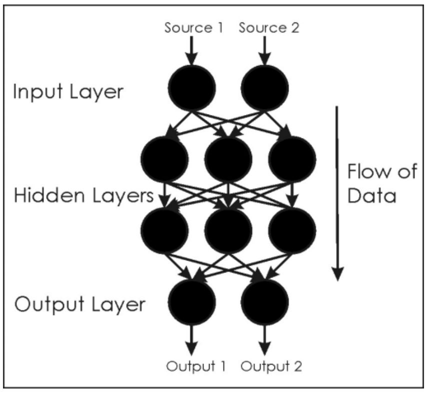

Fig: 1 Neural Network with hidden layers

2.2 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 2.

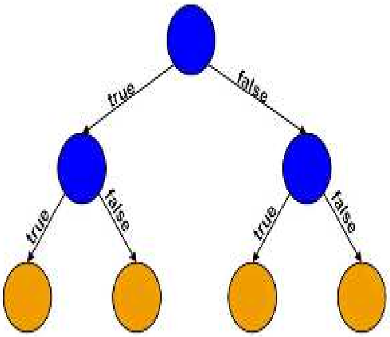

Fig 2 Decision tree

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.

2.3 Genetic Algorithm

Genetic Algorithm attempt to incorporate ideas of natural evaluation The general idea behind GAs is that we can build a better solution if we somehow combine the "good" parts of other solutions (schemata 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. They 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. 3.

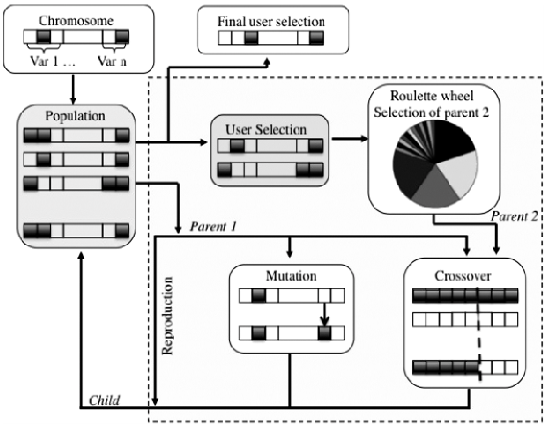

Fig 3: Structural view of Genetic Algorithm

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 a new generation of 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.

2.4 Rule Extraction

The taxonomy of Rule extraction contains three main criteria for evaluation of algorithms: the scope of use, the type of dependency on the black box and the format of the extract description. The first dimension concerns with the scope of use of an algorithm either regression or classification. The second dimension focuses on the extraction algorithm on the underlying black-box: independent versus dependent algorithms. The third criterion focuses on the obtained rules that might be worthwhile: predictive versus descriptive algorithms. Besides this taxonomy the evaluation criteria appears in almost all of these surveys-Quality of the extracted rule; Scalability of the algorithm; consistency of the algorithm [9].

Generally a rule consists of two values. A left hand antecedent and a right hand consequent. An antecedent can have one or multiple conditions which must be true in order for the consequent to be true for a given accuracy whereas a consequent is just a single condition. Thu s while mining a rule from a database antecedent, consequent, accuracy, and coverage are all targeted. Sometimes "interestingness" is also targeted used for ranking. The situation occurs when rules have high coverage and accuracy but deviate from standards. It is also essential to note that even though patterns are produced from rule induction system, they all not necessarily mean that a left hand side ("if "part) should cause the right hand side ("then") part to happen. Once rules are created and interestingness is checked they can be used for predictions in business where each rule performs a prediction keeping a consequent as the target and the accuracy of the rule as the accuracy of the prediction which gives an opportunity for the overall system to improve and perform well.

For data mining domain, the lack of explanation facilities seems to be a serious drawback as it produce opaque model, along with that accuracy is also required. To remove the deficiency of ANN and decision tree, we suggest rule extraction to produce a transparent model along with accuracy. It is becoming increasingly apparent that the absence of an explanation capability in ANN systems limits the realizations of the full potential of such systems, and it is this precise deficiency that the rule extraction process seeks to reduce [10]. Experience from the field of expert systems has shown that an explanation capability is a vital function provided by symbolic AI systems. In particular, the ability to generate even limited explanations is absolutely crucial for user acceptance of such systems. Since the purpose of most data mining systems is to support decision making, the need for explanation facilities in these systems is apparent. But many systems must be regarded as black boxes; i.e. they are opaque to the user.

For the rules to be useful there are two pieces of information that must be supplied as well as the actual rule:

- Accuracy - How often is the rule correct?
- Coverage - How often does the rule apply?

Only because the pattern in the data base is expressed as rule, it does not mean that it is true always. So like data mining algorithms it is equally important to identify and make obvious the uncertainty in the rule. This is called accuracy. The coverage of the rule means how much of the database it "covers" or applies to.

Craven and Shavlik in there paper [11] listed five criteria for rule extraction, and they are as follows:

- **Comprehensibility**: The extent to which extracted representations are humanly comprehensible.
- **Fidelity:** The extent to which extracted representations accurately model the networks from which they were extracted.
- **Accuracy:** The ability of extracted representations to make accurate predictions on previously unseen cases.
- **Scalability:** The ability of the method to scale to networks with large input spaces and large numbers of weighted connections.
- **Generality:** The extent to which the method requires special training.

CONCLUSIONS

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, September2010.

[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.-378-389,1995.

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

[7] AnkitaAgarwal,"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 2005.

[9] Fu Xiuju and Lipo Wang "Rule Extraction from an RBF Classifier Based on Class-Dependent Features ", ISNN'05 Proceedings of the Second international conference on Advances in Neural Networks ,vol.-1,pp.-682-687,2005.

[10] H. Johan, B. Bart and V. Jan, "Using Rule Extraction to Improve the Comprehensibility of Predictive Models". In Open Access publication from Katholieke Universiteit Leuven, pp.1-56, 2006

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

BIOGRAPHIES

Nikita Jain Pursuing M. Tech. in Computer Science She has published many national and international research papers.

Mr. Vishal Shrivastava working as Assistant Professor in Arya College & IT He has published many national and international research papers. He has very depth knowledge of his research areas