International Journal of Research in Advent Technology (IJRAT) Special Issue
E-ISSN: 2321-9637
Available online at www.ijrat.org
National Conference on "Role of Information Technology in Social Innovations"
26th & 27th February 2019

Survey on the Data Mining algorithm based on optimized genetic algorithm and neural network

Rupali R. Deshmukh[1], Prajkta P. Chapke[2]
Assistant Professor, Department of Computer Science and Engineering,
H.V.P.M's College of Engineering and Technology, Amravati, Maharashtra, India.
Email: drupali1604@gmail.com [1], prajkta.chapke@rediffmail.com [2]

Abstract-Now a day's various research are going on in the field of data mining. Which works on volume of data to be stored and processed for query based applications or decision based processes. Data mining refers to the process of extraction of useful information from a pool of data. Various algorithms have been proposed in the past for the mining process out of which neural based mining algorithms are predominant. This paper focuses on work utilization of BPNN integrated with a genetic optimization algorithm for minimizing and bringing about an optimal search value. In this paper the types and existing techniques has been presented.

Index Terms-Algorithm, Data mining, neural network, genetic optimization, KDD knowledge discovery database.

1. INTRODUCTION

With recent advancement in technology and state of art gadgets, the quantity of data and its processing in real time has been growing ever since and the dimensions have increased to such an extent that manual processing and searching of data from a large pool of data is a very tedious and time consuming task. The amount of data is growing very fast and manual analysis, even if possible, cannot keep pace. It cannot be expected that any human can analyze millions of records, each having hundreds of field's thus opening avenues to develop automated techniques for data analysis and processing popularly known as data mining (DM) and Knowledge discovery database (KDD).

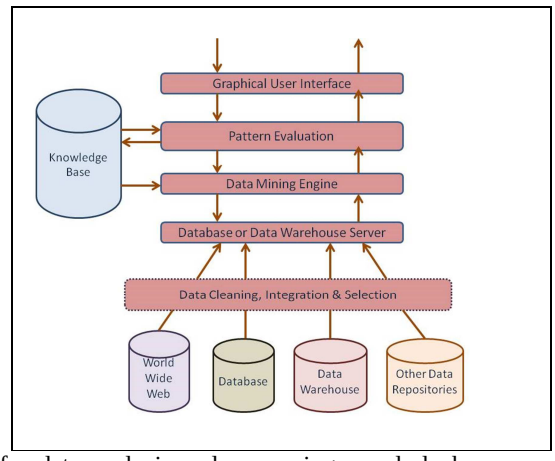


Figure 1. A general data mining system

Figure 1 illustrates a general data mining system consisting of the knowledge base and the data base. Data mining is an iterative process consisting of data cleaning as the first and foremost step. It is similar to pre processing of images and the noisy, redundant and irrelevant data are removed from the data set. This is followed by data integration where the signal from multiple data is combined into a common source. The data most relevant to the query under study is selected in the data selection stage followed by transformation into formats suitable for the mining process. The data sets are subjected to intelligent machine learning techniques or algorithms to extract the patters and the obtained data known as knowledge is presented in a visually interpretable manner.

Data mining (DM) often referred as knowledge discovery in databases (KDD), is a process of nontrivial extraction of implicit, previously unknown and potentiality useful information from a large volume of data [1,2]. The extracted information is also referred as knowledge of the form rules, constraints and regularities. Rule mining is one of critical tasks as they provide a concise statement of potentially important information and most research contributions in the past till data have utilized neural network based techniques for deriving the mining rules. Researchers have been using many techniques such as statistical, AI, decision tree, database, cognitive etc. for rule mining. Several major kinds of data mining methods, including generalization, characterization, classification, clustering, association, evolution, pattern matching, data visualization, and meta-rule guided mining, have been reviewed in [14].

Elaborated discussions on each of the techniques and the algorithms implemented in each of the techniques have also been discussed. Association is one of the best known data mining technique. In association, sequential patterns are discovered based on a relationship between items in the same

transaction [19]. So the association technique is also known as relation technique. The association rule mining technique is used in market basket analysis to identify a set of products that customers frequently purchase together. Frequent item sets play an essential role in many data mining tasks that try to find interesting patterns from databases such as association rules, correlations, sequences, classifiers, clusters and many more of which the mining of association rules is one of the most popular problems [4] [5].

Association rule mining is normally performed in generation of frequent Item sets. The concepts behind association rules are provided at the beginning followed by an overview to some of the previous research works done on this area. On the other hand, sale to predict profit, sale is an independent variable; profit could be a dependent variable. Then based on the past sale and profit data, a regression curve that is used for profit prediction. From the literature, it has been observed that no approaches or tools can guarantee to generate the accurate prediction in the organization. In this paper, they have analyzed the different algorithm and prediction technique. In spite the fact that the least median squares regression is known to produce better results than the classifier linear regression techniques from the given set of attributes. As comparison they found that Linear Regression technique [10]
which takes the lesser time as compared to Least Median Square Regression.

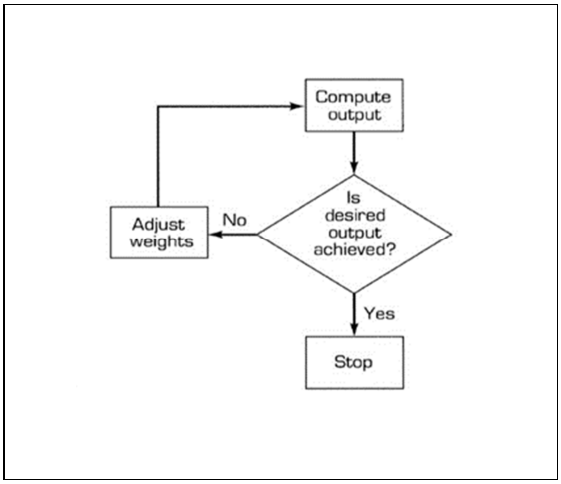
Sequential patterns analysis [9] is one of data mining technique that seeks to discover or identify related patterns, regular events or trends in transaction data over a business period. In sales, with past transaction data, it is easy to identify a set of items that customers buy together in a year. The important heuristics employed includes the optimally sized data structure representations of the sequence database; early pruning of candidate sequences; mechanisms to reduce support counting; and maintaining a narrow search space. Rule mining using neural networks (NNs) is a challenging job as there is no straight way to translate NN weights to rules.

However, NNs have potential to be used in rule mining since they have been found to be a powerful tool to efficiently model data and modeling data is also an essential part of rule mining. As one of branches of DM methods, rule mining aims to apply algorithms of DM to stored data in databases. The core challenge of rule mining research is to turn information expressed in terms of stored data into knowledge expressed in terms of generalized statements about the characteristic of the data which is known as rules. These rules are used to draw conclusions about the whole universe of the dataset. The feed forward NN with unsupervised learning is a good mining tool for discovering data clustering in databases. Discrimination is also an important issue in this regard. Similar patterns should be placed in the same group for discrimination of like patterns in future. In the NN area, Kohonen self-organizing networks [9] or associative memory networks and counter propagation networks all have good potential to be used for this purpose. Data mining is the term used to describe the process of extracting value from a database. A data warehouse is a location where information is stored. The type of data stored depends largely on the type of industry and the company. Classification based techniques [9] [16][18] are used to classify each item in a set of data into one of predefined set of classes or groups. They utilize decision trees, neural network, and statistics. In classification, the authors developed the software that can learn how to classify the data items into groups. Internally the classifiers could use a Bayesian classifier, the conventional support vector machine classifier, back propagation classifier whose methodology and implementation details have been discussed in [16]. The classifier could also be rule based incorporating genetic search technique for the optimal solution search process which forms the motivation behind this work. Artificial neural networks (ANNs) are increasingly used in problem domains involving classification. They are capable at finding commonalities in a set of unrelated data and for this reason are used in a growing number of classification tasks. But there lies, a commonly perceived problem with ANNs when used for classification is that, while a trained ANN [13] can indeed classify the data, sometimes with more accuracy than a traditional, symbolic machine learning approach, the reasons for their classification cannot be found without difficulty. Trained ANNs are commonly perceived to be dark box which map input data onto a class through a number of mathematically weighted connections between neurons. While the idea of ANNs as dark boxes may not be a problem in applications where there is little interest in the reasons behind classification, this can be a major complication in applications where it is vital to have symbolic rules or different forms of knowledge structure, such as identification which are simply interpretable by human experts. From the theoretical perspective, this paper summarizes more literatures and finds the following features.

The genetic algorithm mainly has three merits: First, the genetic algorithm cannot make many mathematical requests to the optimization; second, compared with the local search of tradition, the evolution operation of genetic algorithm enables it to

carry on the effective global optimization; third, the genetic algorithm provided very big flexibility to deal with various domains overlapping issues. The process of the genetic algorithm is mainly: 1st, basis domain of definition and needs the precision to carry on the code; 2nd, initialization population; 3rd, in the sufficiency of individual to the population appraises; 4th, the sufficiency of basis individual, the choice individual carries on the hybrid, variation and some other genetic manipulations, produces the next generation population; 5th, duplicates 2-4 processes, until finding the most appropriate solution.

Hybrid genetic algorithm (HGA) unifies the genetic algorithm in the traditional optimal algorithm the algorithm. To the starting value is quite high as a result of traditional requests of optimal algorithm, the traditional procedure for example the trial and error method, will spend very big time price to obtain the starting value and therefore, the HGA procedure uses the genetic algorithm to obtain the starting value first, then solves this starting value substitution NAND algorithm. Meanwhile, because the genetic algorithm solves the complex constrained optimization problem effect not to be good, will therefore use HGA to solve these issues to receive the quite good effect. The JGA main process is as follows: 1st, basis domain of definition and needs the precision to carry on the code, the encoding method uses the decimal code, the initialization population; 2nd, selecting operation, after only in the choice current population the sufficiency highest individual takes the bee, retains to the population of next generation in; 3rd, after the choice current population the sufficiency highest individual takes the bee, other individuals separately after the bee carries on the interlace operation; 4th, mutation adopts induced mutation operation, not only can guarantee that the multiplicity of population and can the double counting that avoids the blind variation causing; 5th, thought of imitation honeybee evolution, introduces the external population, every other t generations introduce a population, produces through the formula counts n, in the after population of introduction the sufficiency places first n non-bee, the individual substitutes for in the original population the random n non -bees the individual, if introduces population the sufficiency of after bee to be higher than the sufficiency after original population bee, then the after bee of introduction population, after substituting for the original population bee; 6th, after each generation retains bee, had guaranteed the algorithm can restrain the optimal solution, and according to tests to establish the biggest heredity generation number repeatedly.



2. RELATED WORK

Every NN model must be trained with representative data before using. There are basically two types of training, supervised and unsupervised. The basic idea behind training is to pick up set of weights (often randomly), apply the inputs to the NN and check the output with the assigned weights. The computed result is compared to the actual value. The difference is used to update the weights of each layer using the generalized delta rule [7, 11]. This training algorithm is known as 'back propagation'. After several training epochs, when the error between the actual output and the computed output is less than a previously specified value, the NN is considered trained. Once trained, the NN can be used to process new data, classifying them according to its required knowledge. When using supervised training it is important to address the following practical issues.

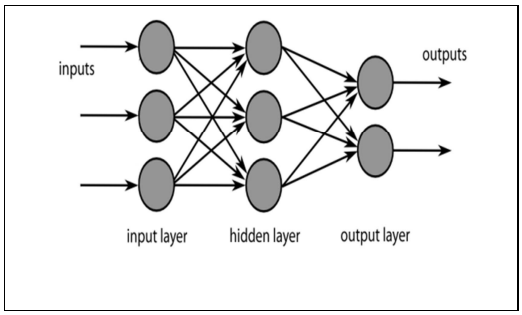
Fig. 2 Weights update flow in proposed work

A. Artificial Neural Network

An Artificial Neuron is basically an engineering approach of biological neuron. It has device with many inputs and one output. ANN is consisting of large number of simple processing elements that are interconnected with each other. In human body work is done with the help of neural network.

Neural Network is just a web of inter connected neurons which are millions and millions in number. With the help of these interconnected neurons all the parallel processing is done in human body and the human body is the best example of Parallel Processing. A neuron is a special biological cell that process information from one neuron to another

neuron with the help of some electrical and chemical change. It is composed of a cell body or soma and two types of out reaching tree like branches: the axon and the dendrites. The cell body has a nucleus that contains information about hereditary traits and plasma that holds the molecular equipments or producing material needed by the neurons. The whole process of receiving and sending signals is done in particular manner like a neuron receives signals from other neuron through dendrites. The Neuron send



signals at spikes of electrical activity through a long thin stand known as an axon and an axon splits this signals through synapse and send it to the other neurons.

The feed-forward neural network architecture is commonly used for supervised learning. Feed - forward neural networks contain a set of layered nodes and weighted connections between nodes in adjacent layers. Feed-forward networks are often trained using a back propagation-learning scheme. Back propagation learning works by making modifications in weight values starting at the output layer then moving backward through the hidden layers of the network.

Figure 3: A multi layered artificial neural

Input: Data set
Target: Classified set

Generate m x n map with a seed neurons
Initialize initial weight W(0)
Select an instance n
Find the winning neuron
Determine the error
Adapt the weight vectors of the k neuron using
genetic optimization after encoding
Repeat steps until convergence
Add or delete connections between neurons
according to the measure distances

B. GA Optimization

The GA is basically based on biological principle of natural selection. The architecture of systems that implement GAs is able to adapt to a wide range of problems. A GA functions by generating a large set of possible solutions to a given problem. It then evaluates each of those solutions, and decides on a fitness level for each solution set. These solutions then breed new solutions. The parent solutions that are more fit are more likely to reproduce, while those that are less fit are more unlikely to do so. In essence, solutions are evolved over time. This way the search space evolves to reach the point of the solution. The GA can be understood by thinking in terms of real life natural selection. The three elements that constitute GAs are the encoding, the operator and the fitness function. The individuals in genetic space are chromosomes. The basic constitution factors are genes. The position of gene in individual is called locus. A set of individuals constructs a population. The fitness represents the evaluation of adaptability of individual to environment.

The elementary operation of genetic algorithm consists of three operands: selection, crossover and mutation. Select is also called copy or reproduction. By calculating the fitness f_i of individuals, it selects high quality individuals with high fitness, copy them to the new population and eliminate the individual with low fitness to generate the new population. Generally used strategies of selection include roulette wheel selection, expectation value selection, paired competition selection and retaining high quality individual selection. Crossover puts individuals in population after selection into match pool and randomly makes individuals in pairs to form

parent generation. Then according to crossover probability and the specified method of crossover, it exchanges part of the genes of individuals that is in pairs to form new pairs of child generation and finally to generate new individuals. Generally used methods of crossover are one point crossover, multi point crossover and average crossover. According to specified mutation rate, mutation substitutes genes with their opposite genes in some loci to generate new individuals.

3. CONCLUSION

A research on data mining based on neural network optimized through genetic algorithm is presented in this paper. Data mining is known for its high robustness, self organizing adaptive, parallel processing capabilities, and distributed storage with a high degree of fault tolerance. The combination of data mining and neural network can greatly improve the efficiency of data mining, and it has been widely used & has presented neural network based data mining scheme to mining classification rules from given databases. An important feature of the rule extraction algorithm is its recursive nature.

Acknowledgments

This section should come before the References. Funding information may also be included here.

Appendix A. Appendix

Appendices should be used only when absolutely necessary. They should come after the References. If there is more than one appendix, number them alphabetically.

REFERENCES

[1] Dr. A.B. Raut, "Neuro-fuzzy, GA-Fuzzy, Neural-Fuzzy-GA: A Data Mining Technique for Optimization ", International Journal of Computer Science and Software Engineering Volume 3, Number 1 (2017), pp. 1-9
[2] M. Charles Arockiaraj "Applications of Neural Networks In Data Mining", International Journal Of Engineering And Science Vol.3, Issue. 1, 2013.
[3] Sachin Sharma and Savita Shiwani, "Data mining based accuracy enhancement of ANN using Swarm intelligence", International journal of communication and computer technologies, Vol. 2, No. 9, 2014.
[4] Bhatia and Jyoti, "An Analysis of heart disease prediction using different data mining techniques, International Journal of Engineering Research and Technology, Vol. 1, pp. 1 – 4, 2012.
[5] Maruthaveni.R, Mrs. Renuka Devi.S.V, "Efficient Data Mining For Mining Classification Using Neural Network", International Journal Of Engineering And Computer Science, Volume. 3, Issue. 2, 2014.
[6] Vidushi Sharma, Sachin Rai, Anurag Dev "A Comprehensive Study of Artificial Neural Networks", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 10, 2012.
[7] Wei Sonalkadu, Prof.Sheetal Dhande "Effective Data Mining Through Neural Network", International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 3, 2012.
[8] Kamruzzaman S M, Jehad Sarkar, "A new data mining scheme using artificial neural networks", Sensors Journal, Vol. 11, No. 5, 2011.
[9] T. Karthikeyan and N. Ravikumar, A Survey on Association Rule Mining International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 1, 2014.
[10] Suguna and Nandhini, "Literature review on data mining techniques",International journal of computer technology and applications, Vol. 6, No. 4, pp. 583 – 585, 2015.
[11] Meenakshi Sharma, "Data Mining: A Literature Survey", International Journal of emerging research in management and technology, Vo. 3, Issue. 2, 2014.
[12] Ranno Agarwal, " Genetic algorithms in data mining", Internaitonal Journal of advanced research in computer science and software engineering, Vol. 5, Issue. 9, 2015.
[13] Kamble, Atul, "Incremental Clustering in Data Mining using Genetic Algorithm", International Journal of Computer Theory and Engineering,Vol. 2, No. 3, 2010.
[14] Shraddha Soni, "A literature review on data mining and its techniques",Indian Journal of applied research, Vol. 5, Issue. 6, 2015.
[15] Tan Jun Shan, He Wei and Qing Yan, "Application of Genetic algorithm in data mining", Proceedings of computer science conference, Vol. 2, pp. 353 – 356, 2009.
[16] Dawei, J. "The Application of Date Mining in Knowledge Management", International Conference on Management of e-Commerce and e-Government, IEEE Computer Society, pp. 7- 9, 2011.

[17] Puneet Chadha and Singh, Classifcation rules and genetic algorithm in data mining", Global journal of comptuer science and technology, software and engineering, Vol. 12, Issue. 15, 2012.
[18] Kantarcıoglu, Murat. Xi, Bowei. Clifton, Chris., "Classifier evaluation and attribute selection against active adversaries", Data
[19] Diti Gupta, Abhishek Singh Chauhan, "Mining Association Rules from Infrequent Item sets: A Survey", International Journal of Innovative Research in Science", Engineering and Technology,Vol.2,Issue 10, 2013.