IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 1, January 2015.
www.ijiset.com
ISSN 2348 – 7968

Review of Soft Computing Methods used in Data Analysis

Mrs. Pranali P. Chaudhari

Department of Information Technology, Savitribai Phule Pune University,
Alandi(D), Pune, Maharashtra, India

Abstract

Due to the huge amount of data that has been processed in various business applications the role of data analysis has become important. Many statistical methods have been proposed that are analyzing data for such applications. But as soon as the size of the data increases the time required and efficiency of these methods reduces. In this paper I have reviewed three soft computing techniques namely ANN, Fuzzy Logic and Genetic Algorithm which have been used in various applications and proved to be better than the traditional statistical based techniques.

Keywords: *ANN, Genetic Algorithm, Fuzzy systems*

1. Introduction

The use of advance computing technologies, business applications and networks requires a huge amount of data to be processed. This data has been collected from various sources and is heterogeneous. Data analysis aims at making use of this data and converts it into information which will help to make decisions. The traditional method of data analysis use statistics but as the amount of data grows or when it comes for real time data processing a need for advanced computational analysis methods arises. Soft computing techniques have proved to be advantageous on efficiently analyzing large data sets.

In this paper three soft computing techniques, Artificial Neural Network, Fuzzy Logic and Genetic Algorithm has been discussed for analyzing huge amount of data.

2. Literature Review

Soft computing is a term applied to a field within computer science which is characterized by the use of inexact solutions to computationally hard tasks for which there is no known algorithm that can compute an exact solution in polynomial time. The main aim of soft computing is to mimic the ability of the human mind to effectively employ modes of reasoning that are approximate rather than exact. The Soft Computing involves three main techniques, the Neural Networks, Genetic Algorithms and fuzzy logic. All of the above mentioned techniques have been used in various applications of different domains and proved to be efficient over the traditional methods used.

Artificial Neural Networks is used as a revenue prediction tool for an enterprise having several malls [1]. The manual calculations and spreadsheets are not sufficient for the calculation due to abundant data and inadequate resources for data analysis. The assumption that all the tenants will extend the lease is not true in real life thus predicting revenue based on this assumption using previous prediction techniques will not work well. Neural networks identify some parameters that affect the revenue income and generate revenue prediction that will help the enterprise in strategic decision making.

Researchers use the Levenberg Marquardt algorithm for training the network. The use of Mean Square Error has been done for evaluating the model. Seven variables are identified for input data: year lease, month lease, the order of the floor, tenant category, the number of units leased per month, leased area and the average monthly rents per square meter per month. The input data consists of 1200 data collected from five years. It is divided into 840 data for training, 180 data for validation and 180 data for testing. The output is measured by MSE and is found that MSE is close to zero at iteration of 191 and is more accurate as compared with the previous method. The researchers show that ANN is 582 times more accurate than previous predictions.

Increasing Honey pot security is another task in which Artificial Neural Networks has proved to be beneficial. The use of ANN is advantageous due to high level of accuracy in real-time operation, low CPU resources utilization during the classification phase and the ability to generalize in order to detect and identify any previously unseen classes [2].

Researcher's uses feed forward neural network together with the Levenberg Marquardt training method. Also the numbers of hidden neurons are set to seven. Fisher's score ranking technique for feature selection is used along with mean and standard deviation which calculate the difference between positive and negative feature. Researchers have created their own data sets consisting of six real malcodes two Trojans (Delf.AAM, Tiny. A), two viruses (W32.Sality.AE, W32.Jeefo) and two worms

ISSN 2348 – 7968

(Autorun, Autorun.MG). The total numbers of features measured are 160. The windows performance tool is used for monitoring the system features and was configured to measure the features every second for 20 minutes for each malcode. Three evaluation measures are used: True Positive Rate (TPR), False Positive Rate (FPR) and Total Accuracy. Thus the use of fisher's score technique for feature selection and the use of ANN model increase the honey pot security.

ANN has also used in weather monitoring systems due to its prediction power [3]. Traditional methods used are based on mathematical and statistical functions, which are very accurate in calculation but not in prediction. The back propagation algorithm is used for predicting the results. The survey was performed on the data collected from udupi district in Karnataka for 8 months of 50 years. The input parameters are the average humidity, and the average wind speed. The output parameters are average rainfall in the 8 months. Out of the total data 70% of data is used for training, 60 samples are kept for validation and remaining 60 for testing. Three techniques: Feed forward with back propagation, Layer recurrent and cascaded feed forward back propagation are used for prediction.
Data analysis in stock market plays a vital role due to the ups and downs happening in the stock market everyday. There are many factors that impact the stock market such as stock index changes in non linear and shares data with high noise characteristics, politics, economy etc. due to these factors the traditional methods are not sufficient. As neural network has an ability to approximate any complex non – linear relation and has robustness and fault tolerance features it is suitable for the stock data analysis. [4]
Researchers have analyzed the stock data from the stock analysis software named "Big Wisdom" which is stored in a file "zong.xls" which is a 4255 x 27 matrix. They have identified some technical indicators and these indicators served as an input to the Back Propagation neural network. The indicators are Moving average (MA) which is used to obtain average cost during a certain period. Random Indicators (KDJ) depends on three lines K, D and J. It not only considers the highest price, lowest price in the calculation period but also takes into account the random amplitude in the course of fluctuation of stock price. Moving Average Convergence/Divergence (MACD) is use to function the signs of aggregation and separation of fast moving average and slow moving average. Relative Strength Index (RSI) compares the average of closing high and the average of closing low. On Balance Volume (OBV) is the degree of the active investors in the stock market. BIAS is the ratio between the application index and the moving average. One output indicator named Increases Scope which is calculated as stock market closing price of today minus the stock market opening price of today divided by the stock market opening price of today.
Researchers have identified 21 input indicators to ensure the accuracy: MA1, MA2, MA3, MA4, MA5, MA6, KDJ.K, KDJ.D, KDJ.J, volume, turnover, MACD.DIFF, MACD.DEA, MACD.MACD, RSI1, RSI2, RSI3, OBV, BIAS1, BIAS2, and BIAS3. The analysis show that RSI1, BIAS1 and volume have larger impact on output which leads the prediction model has a relationship between RSI1, BIAS1, Volume and Increased scope.
For training the network some stocks namely 600839 Sichuan Changhong, 600547 Shandong gold, 000800 faw car and many more are selected and the data is collected from 16 Feb. 2007 to 17 may 2009. The result shows that the predictive growth is very much close to the actual growth.

As discussed earlier the stock data analysis is very important from the view of investors those are investing their money in shares and options. Another technique based on various financial parameters and neuro-fuzzy logic using data mining technique was also introduced. [5] The hybrid technique aims at making the learning algorithm of neural network simpler by combining the linguistic methods and rules of fuzzy logic.
The financial parameters that are considered for share are: Earning per Share (EPS), Dividend per Share (DPS), Price Earning (PE ratio) and Book value. Also the five factors that affect option are Share Price, Exercise Price, and Volatility of the share return, risk free rate of interest and the options time to expiration.
These parameters are given as input to the neuro-fuzzy system and also used to train the network based on the result of these parameters later on data is categorized using the data mining classification technique.

In addition to stock data analysis, exchange rate data analysis neural network has also used in prediction of financial distress for companies. [6] Prediction of financial distress is important because it helps the relevant authorities of company to prevent the occurrences of failure and also for the investor to evaluate and select the company to invest.
Traditional methods used for predicting financial distress include multiple discriminant analysis, logistic regression, multiple regression and soon. Artificial Neural Network was also used in prediction and has proved to be more effective than earlier methods. But ANN also has some limitations such as lack of explanation, lack of feature selection, failing to interpret the classification result etc.
Researchers have concentrated on failing to interpret the classification result drawback of ANN and propose a solution based on group method of data handling technique. GMDH is an iterative method which

ISSN 2348 – 7968

successfully tests models selected from a set of candidate models according to a specified criterion.
The sample data was collected from Taiwan listed companies in the TSE Corporation. Total 200 companies were considered out of which 100 are in financial distress and 100 are healthy. Each case includes 24 financial variables and the GMDH is performed. 80% of data are selected for training, 20% of data is used to validate the accuracy and one year, two year, three year and four year financial distress prediction are constructed. Two types of error type 1 error denoting misclassification rate AND TYPE 2 error denotes all the companies are divided into small, medium and large sized companies. Thus the investor and other peoples can make a correct judgments based upon the GMDH model.

As large amount of information is available on internet the need for new and advance techniques for fast retrieval of information arises. Genetic algorithm has provided an efficient solution for the problem of information retrieval and indexing by grouping different classes of data in an efficient way [7]. Researchers have developed a system for cluster analysis of hypertext documents called SAGH.
SAGH consists of seven modules out of which the first five modules manipulate and transform the documents to construct a normalized matrix. The sixth module perform cluster analysis where the distance of documents from its centroids was calculated and the fitness function through each generation, online and offline performance by generations, the best number of group from generations and evaluation of the population found in each generation is performed. Last module consists of the experimental results.
The documents were obtained from search in Google and are divided in four sets of one hundred documents enclosing the keywords AIDS, cancer, Karl Marx and artificial neural networks. These documents generated the 5 base tests one for each and last with all documents together. These tests aimed at evaluating the coherence of the formed clusters and not the precision. Each time when the results are not satisfactory the chromosomes number is increased and new cluster of higher similarities are formed. For each test some groups are identified that were found accurate. This is the first time when the genetic algorithm is used for hypertext document clustering.

Attracting and retaining users on web sites is the most challenging job for the web site publishers due to the continuous increase in growth and complexity of WWW. Web Usage Mining identifies the patterns that are used to analyze the user's navigational behavior [8]. WUM makes use of user profiles instead of user log files where the details about the URLs representing a user session are stored. In this paper the researchers focused on clustering task of data mining, where the session files are filtered and small sessions are eliminated. But again the removal of this sessions results in loss of information if the number of small sessions are large. As a solution for the problem the fuzzy set theoretic method is proposed in which instead of directly removing the small sessions a weight is assigned to each session using a fuzzy membership function based on the number of URLs accessed by the session. Later on fuzzy c-mean clustering algorithm is applied to discover the cluster of user profiles.
As an input data the web access logs were taken form P.A. College of Engineering, Mangalore website at URL http://www.pace.edu.in. The site hosts variety of information including departments, faculty members, research areas and course information. The logs are recorded for one month. Total 7924 request were logged out of these 6850 are unique URLs. The analysis shows that 78% of URLs are accessed only once, 16% were accessed twice and 6% are accessed three or more time. After calculating access % the maximum access count and average access for each URL is calculated.
For clustering only the URLs that are accessed more than one time are considered. Then the user sessions containing the common URLs are found out. Only the user sessions that contain more than one URL are considered. Thus the researchers conclude that the use of soft computing method is more accurate than the hard computing approach of removing small sessions.
A BP ANN in customer satisfaction is used to identify existing patterns in data and synergies between the attributes of satisfaction [9]. The set of customer satisfaction attributes are used as an input and the output deals with the overall customer satisfaction. The data set was collected by a china company in 2007 collected from 500 customers at the time of interviews. The attributes consists of waiting time, repair time, quality of repair, service attitude, skill of the service worker, technical capability of the product and the reason of fault. The impact factors of all these attributes are calculated separately and then simultaneously. It was observed that reduction in waiting time and repairing time has the highest impact on customer satisfaction. Then comes rising the quality of repairing and service attitude and lastly the skills of the service worker and technical capability of the product. Due to the good predictive capability of ANN it has been proved to be more effective than the traditional regression coefficients and error measures used by statistician.

Analyzing Exchange Data is a complex task. Many decision support systems based on logical reasoning implemented to analyze the exchange rate data. A DSS based on intuitive reasoning expressed by the subjective value on the optimistic – pessimistic axis was proposed in

ISSN 2348 – 7968

2003[10]. Researcher uses this DSS to analyze Yen-Dollar exchange data of 24 hours. This exchange data was filtered based on the dealing range of fluctuation and the trading volume which then used as an input to Neural Network. The Neural Network is trained by these inputs and the certainty factor of subjective evaluation was categorized into 5 parts :- fear(fe), anxiety(an), welter(we), complacency(co) and fever(fr). Also for each evaluation neural network constructed three lands of Neural Network which evaluate from the optimistic-viewpoint (op), moderation-viewpoint (mo) and the pessimistic-viewpoint (pe).
After training the neural network with the trained data the system will evaluate the optimistic-pessimistic condition values and based on those the decision was made. Thus the DSS based on intuitive reasoning has proved to be approximate human thinking.

Artificial Neural Network has used for Indian currency exchange rate forecasting and proved to be better than the other traditional methods based on time series analysis [11].
Researcher uses a multilayer perceptron with two hidden layers and is applied to predict the Indian currency exchange rate. The back propagation algorithm is used for training the neural network
Three indicators previous rate of USD, previous rate of gold and previous rate of crude oil are used as the input to predict the 10 weeks ahead price of USD using neural network. The weekly samples of data from March 31st 1995 to December 31st 2003 are used for the training and samples from Jan 2004 to March 2005 are used for testing purpose.
Both the input and the output data required a preprocessing. These data is normalized in the range -1 to +1. The percentage error between the predicated value and actual value is then calculated and it has been concluded that the average percentage error increases when there is an increase in the block size of data up to some extent. After a particular limit there is not change in data.

A comparative study of using statistical & neural network technique for pattern recognition is highlighted here [12]. Two statistical techniques namely discriminant analysis (DA) and Principal Component Analysis (PCA) and two Neural Network techniques namely Generalized Regression Neural Network and Back Propagation Neural Network are considered. The database considered here is consisting of 400 face images of 40 peoples. Pattern recognition is basically face recognition here. For a given pattern the system has to decide whether it is present in database or not.
Discriminant Analysis is a supervised technique and works by minimizing the intra-class and maximizing the inter-class distances simultaneously. Here the scatter ratio of all the images is computed, for the given input if that scatter ratio lies in any of the group or classes then the image is allotted to that class.
Principal Component Analysis on the other hand is an unsupervised method which calculates the Eigen values and Eigen vectors of the images. For the given images the Eigen value of the test pattern is computed and compared with the Eigen values of all other in the database. If the value matches then the image is present otherwise the image is not present.
Both the neural network techniques create a feature vectors for the training and testing of neural network. The vector was created by taking the mean of image matrix that generated the row vector. This row vector contains mean of all columns. The row vectors of all images in the database are created. 70% of these row vectors are used for training and remaining 30% is used for testing the network.
The comparison of statistical and neural network techniques is made on two parameters: accuracy of recognition and time consumed in recognition. Researchers show that among all the techniques General Regression Neural Network is more efficient.

Pattern recognition is another area where the ANN has been used and produced good results. Here along with the use of neural network for pattern recognition fault tolerance analysis has also been made. Instead of checking for the given pattern as it is, a small distortion is added to the pattern and then it is provided as an input to the system [13]. Fault tolerance analysis is done for both Gaussian and Uniform distribution.
A feed forward back propagation neural network is used. Recognition of pattern is done in two ways: custom recognition (when the precision is high) and force recognition (when the precision is not restricted). Both of these recognition ways are carried out on distorted and original pattern and a comparison is made. Distortion is made by simply changing some bit in the pattern.
Researchers show that for both the recognition methods the recognized pattern is similar to the given pattern. Later a fault tolerance analysis is made for both Gaussian and uniform distribution based on the errors present or encountered at output and hidden layers. It has been shown that fault tolerance capability of Gaussian distribution is more than the uniform distribution.
Document image analysis and recognition is the process of extracting information from the images of documents and interpret the information out of that. There are basically two categories of document image analysis, textual processing and non-textual processing. In textual processing the headline and the paragraph are first separated and then they are linked together according to

ISSN 2348 – 7968

the article [14, 15]. For non-textual processing the components are separated into graphics and photo and different algorithms are applied so that the image display quality is improved.
Authors have used a recursive X-Y cut projection algorithm which recursively based on some threshold values divides the page into different regions. Later the gray scale image is binarized and then it is over segmented into regions. Neural Network is used to classify the regions. A multilayer perceptrons are used and are trained with back propagation algorithm. Finally all the extracted machine generated regions are randomized and separated into training and test sets based on user split rate.
The document images used for training and testing are provided by iArchives. It includes 220 pages from "Chicago Daily Tribune", 124 pages from "The Dallas Morning News", 55 pages from "The Washington Post" and 13 pages from "The Austin American". The small regions are created by the X-Y cut projection algorithm and the test regions are labeled. They are randomly split into training and test set in the ratio of (70% and 30%). The accuracy for all the pages in the data set is then calculated.

Data Preprocessing is the very first step in data mining. Whenever the data is to be analyzed it has to be divided into training and test data sets. Traditional methods for dividing the data sets into training set and test set is based on association rule mining technique [16]. The training set is given as input to ARM technique which derives classification rules and the test set is used to check the accuracy of those rules before validating the rules using a validation set. This traditional method has two main drawbacks first; they don't ensure that training sets fits their test sets with least noise. Second, for an unknown validation set the performance of the rules is not predictable. Thus the researches have proposed a method of splitting data set into sample set and validation set using statistical strategies. Then apply genetic algorithm on this sample set to split into training set and test set. In order to fir the training set and test set with least noise a confusion matrix is prepared and used for computing the confidence factor (CF). Confidence factor for both training and test sets are determined. If CFTraining and CFTest is greater than CFSample and error fit is less than 5% (the specified threshold value) then the fitness function is CFTraining otherwise the fitness function is CFTraining/2.
The data sets considered are obtained from Tom Baker Cancer Center, Canada. The datasets consists of 221 records and 16 attributes. The sample set has 121 and validation set has 100 instances. The average number of generations required to find the best solution, standard deviation and best fitness mean of the solution are then calculated. Thus the splitting of data sets using genetic algorithm is proved to be more accurate.
Genetic Algorithms are widely used for data pre-processing tasks. Feature selection or dimensionality reduction and data reduction are very important data mining tasks. Due to the rapid growth in credit industry there is a need of having data analysis for predicting good and bad applicant group [17]. As an example before giving loan to the applicant the financial party has to perform some analysis based on some characteristics to check whether the loan should be given or not. Since the data set which is to be analyzed is large we should apply feature and instance selection to reduce the data size without compromising on the quality.
Researchers performed five different procedures for constructing the classifier namely: baseline where the classifier is found out without any data preprocessing for comparisons. Feature selection where the genetic algorithm is used to select important feature from the dataset and then this dataset is used to find classifier. Instance selection where based on genetic algorithm the selection of instances is performed on the data set which then used to construct the classifier. Lastly there was feature selection + instance selection and instance selection + feature selection where both are applied on the dataset one after the another as mentioned and then the classifier is developed. To perform instance and feature selection genetic algorithm make use of three parameters: population size, crossover rate and mutation rate.
The whole procedure is performed on the German Credit dataset containing 1000 data sample out of which 700 are good and 300 are bad cases. Also 20 different features are identified which describes each data sample. Finally it s concluded that the genetic algorithm can obtain better results as compared to SVM and K-NN in terms of better instance selection rates and higher classification accuracy.

3. Conclusions

Due to the increase in size of data and variety of data sets, the need of data analysis has also increased. But to select an appropriate algorithm for data analysis is a complex task because it requires considering various parameter like time, efficiency, accuracy, fault tolerance etc. Soft computing techniques are the alternative to the traditional statistical technique and are meant for handling large variety of data.
In this paper I have reviewed various soft computing technique used in different domains and their advantages over the traditional techniques. Also the techniques such as Bayesian network, Back propagation algorithm, feed forward network, and genetic algorithms are found to be useful for solving the problems during data analysis.

ISSN 2348 – 7968

The task of selecting correct attribute for learning algorithm is very challenging job because it directly affects the accuracy of the system. As far as exchange rate forecasting is considered the parameters like previous rate of USD, previous rate of gold and previous rate of crude oil are prove to be more accurate than the parameter based on statistical metrics like Normalized Mean Square Error (NMSE), Mean Absolute Error (MAE) and Directional Symmetry (DS). For stock option data analysis, the use of financial parameters like Earning Per Share (EPS), Dividend Per Share (DPS), Price Earning, PE Ratio and Book Value have given good result as compare to statistical based parameter like Moving Average (MA), Relative Strength Index (RSI), On Balance Volume (OBV) and many more, even when same learning algorithm is used but there is still some scope to research in the method based on pattern learning for financial prediction.

References

[1] Christine Sanjaya, May Liana, Agus Widodo, "Revenue Prediction using Artificial Neural Network, IEEE International Conference on Advances in Computing, Control, and Telecommunication Techniqiues, 2010.
[2] Milad Daliran, Ramin Nassiri, Golamreza Latif-Shabgahi, "Using Data Analyis by deploying Artificial Neural Network to increase Honeypot Security", IEEE, 2008.
[3] Kumar Abhishek, Abhay Kumar, Rajeev Ranjan, Sarthak Kumar, "A rainfall prediction model using Artificial neural network", IEEE Control and System Graduate Research Colloquium, 2012.
[4] Zhou Yixin, Jie Zhang, "Stock Data Analysis based on BP neural network" IEEE International Conference on Communication Software and Networks, 2010.
[5] Anupama Surendran, "Data Mining Technique to analyze the risks in stocks/options Investment", IEEE, 2009.
[6] Chien-hui yang, Mou-yuan liao, Pin-lun chen, Mei-ting huang, Chun-wei Huang, Jia-Siang Huang, Jui-bin Chung, "Constructing Financial Distress prediction model using group method of data handling technique", IEEE International Conference on Machine Learning and Cybernetics, Baoding, 2009.
[7] Lando M. di Carlantonio, Rosa Maria E. M. da Costa, "A genetic system for cluster analysis for hypertext documents", IEEE International Conference on Intelligent Systems Design and Applications, 2007.
[8] Zahid Ansari, A. Vinaya Babu, Waseem Ahmed, Mohammad Fazle Azeem, "A fuzzy set theoretic approach to discover user sessions from web navigational data", IEEE, 2011.
[9] Guozheng Zhang, Faming Zhou, Junfeng Liu, Yong Lan, "Customer satisfaction data analysis based on BP ANN", IEEE, 2008.
[10] Kyuichiro Tani, Katsuari Kamel, "An Exchange Data Analysis Support System by intuitive reasoning based on Neural network", IEEE, 2003.
[11] S. S. Gill, Amanjot Kaur Gill, Naveen Goel, "Indian Currency Exchange Rate forecasting using Neural Network", IEEE, 2010.
[12] Tasweer Ahmad, Ahlam Jameel, Dr. Balal Ahmad, "Pattern Recognition using Statistical & Neural Techniques", IEEE, 2011.
[13] Patavardhan Prashant, D.H. Rao, Anita G. Deshpande, " Fault tolerance Analysis of Neural Network for Pattern Recognition", IEEE International Conference on Computational Intelligence and Multimedia Applications, 2007.
[14] Wei Zhang, Timothy L. Andersen, "Using Artificial Neural Networks to Identify headings in Newspaper Documents", IEEE, 2003.
[15] Tim Andersen, Wei Zhang, "Features for Neural Net Based Region identification of newspaper documents", IEEE, 2003.
[16] Janaki Gopalan, Erkan Korkmaz, Reda Alhajj, Ken Barker, "Effective Data Mining by Integrating Genetic Algorithm into the Data Preprocessing Phase", IEEE, 2005.
[17] Chih –Fong Tsai, Jui-Sheng Chou, "Data Pre-processing by Genetic Algorithms for Bankruptcy Prediction", IEEE, 2011.