Efficient Data Mining for proper Mining Classification using Neural Networks

A.Kalaivani

Assistant Professor
Department of Information Technology
Sri GVG Visalakshi College for Women
Udumalpet,Tamil Nadu,India

***Abstract*— With the development of database, the data volume stored in database increases rapidly and in the large amounts of data much important information is hidden. If the information can be extracted from the database they will create a lot of profit for the organization. The question they are asking is how to extract this value. The answer is data mining. There are many technologies available to data mining practitioners, including Artificial Neural Networks, Genetics, Fuzzy logic and Decision Trees. Many practitioners are wary of Neural Networks due to their black box nature, even though they have proven themselves in many situations. This paper is an overview of artificial neural networks andquestions their position as a preferred tool by data mining practitioners.**

***IndexTerms*—ANN- Artificial Neural Networks, ESRNN- Extraction of Symbolic Rules from ANN's, data mining, symbolic rules.**

I. INTRODUCTION

Data mining is the term used to describe the process of extracting value from a database. A datawarehouse is a location where information is stored. The type of data stored depends largely on the type of industry and the company. Following example of a financial institution failing to utilize their datawarehouse. Income is a very important socio-economic indicator. If a bank knows a person's income, they can offer a higher credit card limit or determine if they are likely to want information on a home loan or managed investments. Even though this financial institution had the ability to determine a customer's income in two ways, from their credit card application, or through regular direct deposits into their bank account, they did not extract and utilize this information [1,2].

An artificial neural network (ANN), usually called neural network (NN), is a mathematical model or computational model that is inspired by the structure or functional aspects of biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation. ANN is an adaptive system that changes its structure based on external or internal information that flows through the networkduring the learning phase. They are used to model complex relationships between inputs and outputs or to find patterns in data. Example Facial or Handwriting or Voice Recognition [3].

In this paper we discuss a data mining scheme, referred to as ESRNN (Extraction of Symbolic Rules from ANNs) to extract symbolic rules from trained ANNs. A three-phase training algorithm. In the first and second phases, appropriate network architecture is determined using weight freezing based constructive and pruning algorithms. In the third phase, symbolic rules are extracted using the frequently occurred pattern based rule extraction algorithm by examining the activation values of the hidden nodes [10].

II. INTRODUCTION OF DATA MINING

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. Example of a financial institution failing to utilize their data-warehouse is in cross-selling insurance products (e.g. home, life and motor vehicle insurance). By using transaction information they may have the ability to determine if a customer is making payments to another insurance broker. This would enable the institution to select prospects for their insurance products.[1,2]

Need of Data Mining

Finding information hidden in data is as theoretically difficult as it is practically important. With the objective of discovering unknown patterns from data, Companies have been collecting data for decades, building massive data warehouses in which to store it. Even though this data is available, very few companies have been able to realize the actual value stored in it. The question these companies are asking is how to extract this value. The answer is Data mining [1,2]

Techniques/Functionalities of Data Mining

There are two fundamental goals of data mining: prediction and description. Prediction makes use of existing variables in the database in order to predict unknown or future values of interest, and description focuses on finding properties that describe the existing data.[3].There are several data mining techniques fulfilling these objectives. Some of these are associations, classifications, sequential patterns and clustering. Another approach of the study of data mining techniques is to classify the techniques as: userguided or verification-driven data mining and, discovery-driven or automatic discovery of rules.

A.Association Rules :

An association rule is an expression of the form X => Y, where X and Y are the sets of items. The meaning of such a rule is that the transaction of the database, which contains X tends to contain Y. Given a database, the goal is to discover all the rules that have the support and confidence greater than or equal to the minimum support and confidence, respectively.

Support means how often X and Y occur together as a percentage of the total transactions. Confidence measures how much a particular item is dependent on another. Patterns with a combination of intermediate values of confidence and support provide the user with interesting and previously unknown information.

B. Classification Rules:

Classification involves finding rules that partition the data into disjoint groups. The input for the classification data set is the training data set, whose class labels are already known. Classification analyses the training data set and constructs a model based on the class label, and aims to assign class label to the future unlabelled records. Since the class field is known, this type of classification is known as supervised learning. There are several classification discovery models. They are: the decision tree, neural networks, genetic algorithms and some statistical models.

C. Clustering

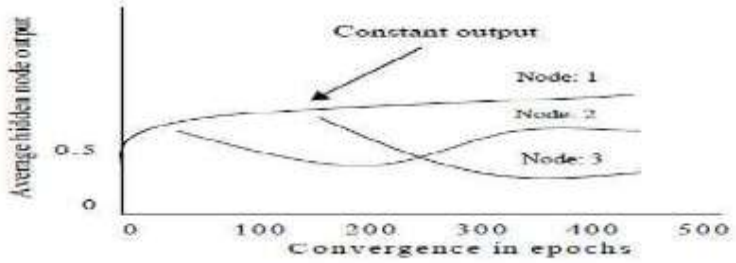
Clustering is a method of grouping data into different groups, so that the data in each group share similar trends and patterns. The goal of the process is to identify all sets of similar examples in the data, in some optimal fashion If a measure of similarity is available, then there are a number of techniques for forming clusters. It is an Unsupervised classification.

Heuristic Clustering Algorithm[10]

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar within the same cluster and are dissimilar to the objects in other clusters. A cluster of a data objects can be treated collectively as one group in many applications. There exist a large number of clustering algorithms, such as, k-means, kmenoids. The choice of clustering algorithm depends both on the type of data available and on the particular purpose and applications.

After applying pruning algorithm in ESRNN, the ANN architecture produced by the weight freezing based constructive algorithm contains only important nodes and connections. Therefore, rules are not readily extractable because the hidden node activation values are continuous. The separation of these values paves the way for rule extraction. It is found that some hidden nodes of an ANN maintain almost constant output while other nodes change continuously during the whole training process Figure shows output of three hidden nodes where a hidden node maintains almost constant output value after some training epochs but output value of other nodes are changing continually.

Fig. 1



In ESRNN, no clustering algorithm is used when hidden nodes maintain almost constant output value. If the outputs of hidden nodes do not maintain constant value, a heuristic clustering algorithm is used.

The aim of the clustering algorithm is to separate the output values of the hidden nodes. Consider that the number of hidden nodes in the pruned network is H. Clustering the activation values of the hidden node is accomplished by a simple greedy algorithm that can be summarized as follows:

1. Find the smallest positive integer d such that if all the network activation values are rounded to d decimal places, the network still retains its accuracy rate
2. Represent each activation value α by the integer closest to α × 10d. Let Hi = <hi,1, hi,2, .., hi,k> be the k-dimensional vector of these representations at hidden node i for patterns x1, x2 , . . . , xk and let H = (H1, H2, . . . , HH) be the k × H matrix of the hidden representations of patterns at all H hidden nodes.
3. Let P be a permutation of the set {1, 2, . . . , H} and set m = 1.
4. Set i = P(m).
5. Sort the values of the ith column (Hi) of matrix H in increasing order.
6. Find a pair of distinct adjacent values hi,j and hi, j+1 in Hi such that if hi, j+1 is replaced by hi,j no conflicting data will be generated.
7. If such a pair of values exists, replace all occurrences of i, j 1 h + in Hi by i, j h and repeat Step 6. Otherwise, set m = m+1. If m ≤ H, go to Step 4, else stop.

The activation value of an input pattern at hidden node m is computed as the hyperbolic tangent function, it will have a value in the range of [−1, 1]. Steps 1 and 2 of the clustering algorithm find integer representations of all hidden node activation values. A small value for d in step 1 indicates that relatively few distinct values for the activation values are sufficient for the network to maintain its accuracy.

The array P contains the sequence in which the hidden nodes of the network are to be considered. Different ordering sequences usually result in different clusters of activation values. Once a hidden node is selected for clustering, the separated activation values are sorted in step 5 such that the activation values are in increasing order. The values are clustered based on their distance. We implemented step 6 of the algorithm by first finding a pair of adjacent distinct values with the shortest distance. If these two values can be merged without introducing conflicting data, they will be merged. Otherwise, a pair with the second shortest distance will be considered. This process is repeated until there are no more pairs of values that can be merged.

The next hidden node as determined by the array P will then be considered.

III. INTRODUCTION OF NEURAL NETWORKS

An Artificial Neuron is basically an engineering approach of biological neuron. It has device with many inputs and one output. ANN is consist of large number of simple processing elements that are interconnected with each other and layered also 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 this interconnected neurons all the parallel processing is done in human body and the human body is the best example of Parallel Processing. Example Facial or Handwriting or Voice Recognition[6] 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 receive 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.[6]

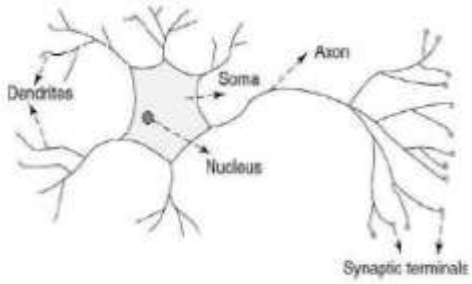


Fig 2 Human Neurons

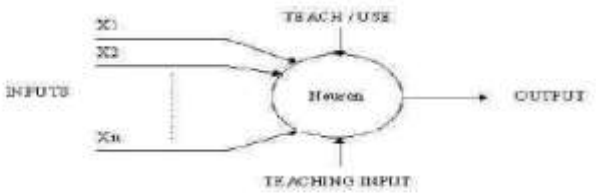


Fig 3 Artificial Neuron

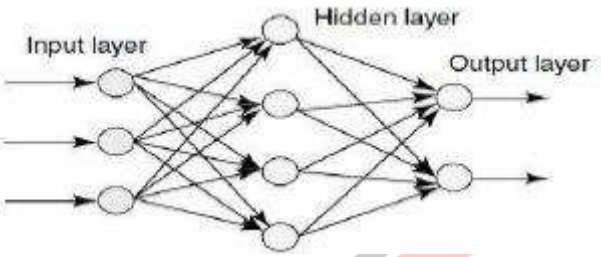


Fig 4 Multilayered ANN

IV. IMPLEMENTATION OF NEURAL NETWORKS IN DATA MINING

Effective Combination of Neural Network and Data Mining Technology:

The technology almost uses the original ANN software package or transformed from existing ANN development tools, the workflow of data mining should be understood in depth, the data model and application interfaces should be described with standardized form, then the two technologies can be effectively integrated and together complete data mining tasks. Therefore, the approach of organically combining the ANN and data mining technologies should be found to improve and optimize the data mining technology.[4]

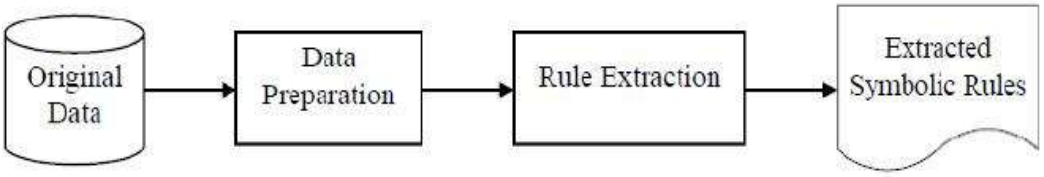


Figure 5. Data mining technique using ANNs.[10,11]

The planned data processing theme consists of two steps: data preparation and rule extraction.

Data Preparation

One must prepare quality information by pre-processing the data. The input to the data mining algorithms is assumed to be distributed, containing incorrect values or no missing wherever all options square measure vital. The real-world data could also be noisy, incomplete, and inconsistent, which might disguise helpful patterns. data preparation could be a method of the first information to form it acceptable a particular data mining technique. The data mining using ANNs can only handle numerical data. There are different kinds of attributes that must be representing input and output attributes.

Real-valued attributes square measure sometimes rescaled by some function that maps the value into the range 0…1 or −1…1

Integer-valued attributes square measure most often handled as if they were real-valued. If the amount of various values is only

small, one among the representations used for ordinal attributes may additionally be applicable.
Ordinal attributes with m different prices are either mapped onto an equidistant scale creating them pseudo-real-valued or are represented by m −1 inputs of that the leftmost k have value 1 to represent the k-th attribute value whereas all others are 0.

V. CONCLUSION

In this paper, We present research on data mining based on neural network. At present, data mining is a new and important area of research, and neural network itself is very suitable for solving the problems of data mining because its characteristics of good robustness, self-organizing adaptive, parallel processing, distributed storage, high degree of fault tolerance & network structure The combination of data mining and neural network can greatly improve the efficiency of data mining, and it has been widely used & we have presented neural network based data mining scheme to mining classification rules from given databases. This work is an attempt

to apply the approach to data mining by extracting symbolic rules. An important feature of the rule extraction algorithm is its recursive nature. A set of experiments was conducted to test the approach using a well defined set of data mining problems. The results indicate that, using the approach, high quality rules can be discovered from the given data sets. The extracted rules are concise, comprehensible, order insensitive, and do not involve any weight values. The accuracy of the rules from the pruned network is as high as the accuracy of the fully connected networks. Experiments showed that this method helped a lot to reduce the number of rules significantly without sacrificing classification accuracy. In almost all cases ESRNN outperformed the others. With the rules extracted by the method here, ANNs should no longer be regarded as black boxes. Since, black boxes are diminished & more researchers use them. Thus, neural networks are becoming very popular with data mining practitioner.

REFERENCES

[1] M.Charles Arockiaraj “Applications of Neural Networks In Data Mining”, Arakkonam, (Research Inventy: International Journal Of Engineering And Science Vol.3, Issue1),May 2013.

[2] Dr. Yashpal Singh ,Alok Singh Chauhan “Neural Networks In Data Mining” , India , (Journal of Theoretical and Applied Information Technology)2005.

[3] K. Amarendra, K.V. Lakshmi & K.V. Ramani “Research on Data Mining Using Neural Networks” , India

[4] Xianjun Ni “Research of Data Mining based on Neural Networks” ,China , (World Academy of Science, Engineering and Technology Vol:2) ,2008.

[5] Sonalkadu, Prof.Sheetal Dhande “Effective Data Mining Through Neural Network”, (International Journal of Advanced Research in Computer Science and SoftwareEngineering Volume 2, Issue 3) ,March 2012

[6] Vidushi Sharma ,Sachin Rai ,Anurag Dev “A Comprehensive Study of Artificial Neural Networks”, India (International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 10) ,October 2012

[7] Ms. Sonali. B. Maind ,Ms. Priyanka Wankar “Research Paper on Basic of Artificial Neural Network”, Wardha ,(International Journal on Recent and Innovation Trends in Computing and Communication Volume: 2 Issue: 1),January 2014.

[8] Ani1 K. Jain ,Jianchang Mao ,K.M. Mohiuddin “Artificial Neural Networks : A Tutorial” , Michigan ,March 1996

[9] Ajith Abraham “Artificial Neural Networks” Oklahoma State University, Stillwater, USA 2005.

[10] S. M. Kamruzzaman and A. M. Jehad Sarkar “A New Data Mining Scheme Using Artificial Neural Networks”, Korea , 28 April 2011.

[11] Mrs.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) February , 2014.