PERFORMANCE ANALYSIS OF DATA MINING ALGORITHMS WITH NEURAL NETWORK

Ms. Aruna J. Chamatkar
Research Scholar, Department of Electronics & Computer Science,
RTM Nagpur University, Nagpur, India

Dr. P.K. Butey
Research Supervisor, HOD Computer Science Department,
Kamla Nehru Mahavidyalaya, Nagpur India

ABSTRACT

Data mining has evolved into an active and important area of research because of previously unknown and interesting knowledge from very large real-world database. Data mining methods have been successfully applied in a wide range of unsupervised and supervised learning applications. Classification is one of the data mining problems receiving great attention recently in the database community. Neural networks have not been thought suited for data mining because how the classifications were made is not explicitly stated as symbolic rules that are suitable for verification or interpretation by humans. The neural network is first trained to achieve the required accuracy in data mining. A network pruning algorithm is used to remove redundant connections of the network. The activation values of the hidden units in the network are analyzed in the network and classification rules are generated using the result of this analysis. In this paper we are going to combine neural network with the three different algorithms which are commonly used in data mining to improve the data mining result. These three algorithms are CHARM Algorithm, Top K Rules mining and CM SPAM Algorithm.

Keywords: Artificial Neural Network, Data Mining, CHARM algorithm, CM -SPAM Algorithm, K -rule mining.
I. INTRODUCTION

In present day human beings are used in the different technologies to adequate in there society. Every day the human beings are using the vast data and these data are in the different fields. It may be in the form of documents, may be graphical formats, may be the video, may be records (varying array). As the data are available in the different formats so that the proper action to be taken for better utilization of the available data. As and when the customer will require the data should be retrieved from the database and make the better decision.

This technique is actually we called as a data mining or Knowledge Hub or simply KDD (Knowledge Discovery Process). The important reason that attracted a great deal of attention in information technology the discovery of useful information from large collections of data industry towards field of “Data mining” is due to the perception of “we are data rich but information poor”. Very huge amount of data is available but we hardly able to turn them in to useful information and knowledge for managerial decision making in different fields. To produce information it requires very huge database. It may be available in different formats like audio/video, numbers, text, figures, and Hypertext formats. To take complete advantage of data; the data retrieval is simply not enough, it requires a tool for extraction of the essence of information stored, automatic summarization of data and the discovery of patterns in raw data.

With the enormous amount of data stored in databases, files, and other repositories, it is very important to develop powerful software or tool for analysis and interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. The only answer to all above is ‘Data Mining’. Data mining is the method of extracting hidden predictive information from large databases; it is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses [1][2][3][4]. Data mining tools predict behaviors and future trends help organizations and firms to make proactive knowledge-driven decisions [2]. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by prospective tools typical of decision support systems. Data mining tools can answer the questions that traditionally were too time consuming to resolve. They created databases for finding predictive information, finding hidden patterns that experts may miss because it lies outside their expectations.

One of the data mining problems is classification. Various classification algorithms have been designed to tackle the problem by researchers in different fields such as mathematical programming, machine learning, and statistics. Recently, there is a surge of data mining research in the database community. The classification problem is re-examined in the context of large databases. Unlike researchers in other fields, database researchers pay more attention to the issues related to the volume of data. They are also concerned with the effective use of the available database techniques, such as efficient data retrieval mechanisms. With such concerns, most algorithms proposed are basically based on decision trees.

II. ARTIFICIAL NEURAL NETWORK

An artificial neural network (ANN), often just called a "neural network", is a computational model or mathematical model based on biological neural networks, in other words we can say that it is an emulation of biological neural system. Artificial neural network consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. Most of the time ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase [8].
A. Training Of Artificial Neural Networks

A neural network has to be configured such that the application of a set of inputs produces (either ‘direct’ or via a relaxation process) the desired set of outputs. Different methods to set the strengths of the connections presents. Prior knowledge is a one way is to set the weights explicitly. Another way is to ‘train’ the neural network by feeding it teaching patterns and letting it change its weights according to some learning rule. We can categorize the learning situations as follows:

- Supervised learning or Associative learning in which the network is trained by providing it with input and matching output patterns. These pairs of input-output can be provided by the system which contains the neural network (self-supervised) or by an external teacher.
- Unsupervised learning or Self-organization in which an (output) unit is trained to respond to clusters of pattern within the input. The system is supposed to discover statistically salient features of the input population in this paradigm. Unlike the supervised learning paradigm, there is no other priori set of categories into which the patterns are to be classified; rather the system must develop its own representation of the input stimuli.
- Reinforcement learning this type of learning may be considered as an intermediate form of the above two learning types. Here the learning machine does some action on the environment and gets a feedback response from the environment system. The learning system grades its action bad (punishable) or good (rewarding) based on the environmental response and accordingly adjusts its parameters.

In more practical terms neural networks are non-linear statistical data modeling tools. They can be used to model complex relationships between inputs and outputs or to find patterns in data. Using neural networks as a tool, data warehousing firms are harvesting information from datasets in the process known as data mining. The difference between these data warehouses and ordinary databases
is that there is actual manipulation and cross-fertilization of the data helping users makes more informed decisions.

Neural networks essentially comprise three pieces: the architecture or model; the learning algorithm; and the activation functions. Neural networks are programmed or “trained” to “store, recognize, and associatively retrieve patterns or database entries; to solve combinatorial optimization problems; to filter noise from measurement data; to control ill-defined problems. It is precisely these two abilities (pattern recognition and function estimation) which make artificial neural networks (ANN) so prevalent a utility in data mining system. As a size of data sets grows massive, the need for automated processing becomes clear. With their “model-free” estimators and their dual nature, neural networks serve data mining in a myriad of ways.

Data mining is the business of answering questions that you’ve not asked yet. Data mining reaches deep into databases. Data mining tasks can be classified into two categories: Descriptive and predictive data mining. Descriptive data mining provides information to understand what is happening inside the data without a predetermined idea. Predictive data mining allows the user to submit records with unknown field values, and based on previous patterns discovered from the database system will guess the unknown values. Data mining models can be categorized according to the tasks they perform: Prediction and Classification, Association Rules, Clustering. Prediction and classification is a predictive model, but clustering and association rules are descriptive models.

Figure 2: Neural-based framework for ARM.
The most common action in data mining is classification. Classification recognizes patterns that describe the group to which an item relates. It does this by examining existing items that already have been classified and inferring a set of rules. The major difference being that no groups have been predefined. Prediction is the construction and use of a model to assess the class of an unlabeled object or to assess the value or value ranges of a given object is likely to have. The next application is forecasting. This is different from predictions because it estimates the future value of continuous variables based on patterns within the database. Neural network works depending on the provide associations, classifications, architecture, clusters, forecasting and prediction to the data mining industry.

B. Feedforward Neural Network

A feedforward neural network (FFNN) is the one of the simplest neural network technique, such as in Figure, consists of three layers: an input, hidden and output layer. In every layer there are one or more processing elements (PEs). PEs is meant to simulate the neurons in the brain and this is why they are often referred to as nodes or neurons. A processing element receives inputs from either the previous layer or the outside world. There are connections between the processing elements in each layer that have a weight (parameter) associated with the element. During training this weight is adjusted. Information only travels in the forward direction through the network - there are no feedback loops.

![Figure 3: Flow of the feedforward neural network](image)

The simplified process for training a feedforward neural network is as follows:

1. Input data is presented to the network and propagated through the network until it reaches the output layer of the network. This forward process produces a predicted output.
2. The predicted output is subtracted from the actual output and an error value for the networks is calculated.
3. The neural network then uses supervised learning, most of the cases which is back propagation for training the network. Back propagation is a learning algorithm for adjusting the weights. It starts with the weights between the output layer PE’s and the last hidden layer PE’s and works backwards through the network.
4. The forward process starts again once back propagation has finished, and this cycle is continued until the error between predicted and actual outputs is minimized.
C. The Back Propagation Algorithm

Propagation of error or Back propagation is a common method of teaching artificial neural networks how to perform a particular task. The back propagation algorithm is used in layered feedforward ANNs. This means that the artificial neurons are organized in layers, and send their signals “forward”, and then the errors are propagated backwards. The back propagation algorithm uses supervised learning, which means that we provide the algorithm with examples of the inputs and outputs we want the network to compute, and then the error (difference between actual and expected results) is calculated. The idea of the back propagation algorithm is to reduce this error, until the ANN learns the training data.

Algorithm:
1. Initialize the weights in the network (often randomly)
2. repeat
   * for each example e in the training set do
     1. O = neural-net-output(network, e) ; forward pass
     2. T = teacher output for e
     3. Calculate error (T - O) at the output units
     4. Compute delta_wi for all weights from hidden layer to output layer ; backward pass
     5. Compute delta_wi for all weights from input layer to hidden layer ; backward pass continued
     6. Update the weights in the network
   * end
3. until all examples classified correctly or stopping criterion satisfied
4. return(network)

III. DIFFRENT DATA MINING ALGORITHMS WITH NEURAL NETWORK

A. Association rule mining with neural network

In this paper we are going to implement the neural network with the different data mining techniques we discussed in the previous review paper we published.[9] The flowchart of the proposed Association rule mining with neural network shows in figure 4 . Firstly the Trained Dataset with the known input and known output is given to the system as the input. The discretization is applied on the input to obtain the association rules from the dataset. Classification rule set is obtained from the discretization. Classification Rule Neural Network (CRNN) construction and parameter learning set is then obtain from the classification rule set. This constructor and learning parameter is then used to form the classification model of the CRNN. Now we apply the testing dataset with the unknown output as the input to the classification model of the CRNN to find the known rules from the input. The output of the CRNN classification is the desired output dataset.
Figure 4: Association rule mining with neural network

A. CHARM Algorithm with Neural Network
CHARM is an efficient algorithm for enumerating the set of all frequent closed item-sets. There are a number of innovative ideas employed in the development of CHARM; these include:

1) CHARM simultaneously explores both the item-set space and transaction space, over a novel IT-tree (item set-tides tree) search space of the database. In contrast previous algorithms exploit only the item-set search space. The integration of the CHARM algorithm with the neural network is shown in figure 5.

2) CHARM uses a highly efficient hybrid search method that skips many levels of the IT-tree to quickly identify the frequent closed item-sets, instead of having to enumerate many possible subsets.

3) It uses a fast hash-based approach to eliminate non-closed item-sets during subsumption checking. CHARM also able to utilize a novel vertical data representation called diffset [5], for fast frequency computations. Diffsets also keep track for differences in the tids of a candidate pattern from its prefix pattern. Diffsets drastically cut down (by orders of magnitude) the size of
memory required to store intermediate results. Thus the entire working set of patterns can fit entirely in the memory, even for huge databases.

Figure 5: CHARM algorithms with neural network

B. CM SPAM Algorithm with Neural Network

Mining useful patterns in sequential data is a challenging task. Many studies have been proposed for mining interesting patterns in sequence databases [6]. Sequential pattern mining is probably the most popular research topic among them. A subsequence is called sequential pattern or frequent sequence if it frequently appears in a sequence database and its frequency is no less than a user-specified minimum support threshold minsup [7]. Sequential pattern mining plays an important role in data mining and is essential to a wide range of applications such as the analysis of web medical data, program executions, click-streams, e-learning data and biological data [6]. Several efficient algorithms have been proposed for sequential data mining and one of them is CM SPAM Algorithm. The integration of the CM SPAM algorithm from the sequential data mining and neural network is shown in figure 6. First the trained dataset with known input and output is given as the input to the system. Discretization is done if it necessary it gives the sequential classification from the known input and sequential classification form the known output. Combination of this both dataset gives us the
classification sequential dataset. Then CM SPAM algorithm is applied to get the sequential construction, parameter learning and optimized dataset from the sequential dataset. Combination of all this parameters gives us the Classification model of Sequential CM SPAM, which is then used on the testing dataset with unknown output for the sequential data mining.

![Diagram](image)

**Figure 6:** CM SPAM algorithms with neural network

### IV. EXPERIMENTAL RESULT

Once the algorithms were implemented and the results were summarized, we tried to optimize the algorithms with the help of Neural Network techniques. Out of the many techniques available, we selected Back Propagation and FeedForward with Back Propagation. Each of the networks when applied to each of the Data Mining algorithms gave some interesting and conclusive results which we analyzed to find the best possible application for the given combination of dataset.

In all the techniques, the neural network was first trained with the input dataset of 500 entries, and the output dataset of mined values. Once the network was trained, we used the network for mining of 1000 and 2000 records. The same process was followed for each of the neural networks and corresponding weights were evaluated for each network. Once the networks are trained, we evaluated each of the networks with a different length of input dataset in order to get the desired data mining results. The mining results obtained are shown as follows,
Table 1: Top K Rules Mining (Mining top 10% Rules) with different neural network techniques

<table>
<thead>
<tr>
<th></th>
<th>Top K Rules</th>
<th>BPP</th>
<th>FF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time 1000</td>
<td>24</td>
<td>31</td>
<td>7</td>
</tr>
<tr>
<td>Memory 1000</td>
<td>8.28</td>
<td>5.01</td>
<td>2.83</td>
</tr>
<tr>
<td>Computational</td>
<td>-----</td>
<td>32%</td>
<td>13%</td>
</tr>
<tr>
<td>Complexity 1000</td>
<td>-----</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time 2000</td>
<td>171</td>
<td>119</td>
<td>58</td>
</tr>
<tr>
<td>Memory 2000</td>
<td>15.489</td>
<td>7.62</td>
<td>4.72</td>
</tr>
<tr>
<td>Computational</td>
<td>-----</td>
<td>31%</td>
<td>14%</td>
</tr>
<tr>
<td>Complexity 2000</td>
<td>-----</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: CHARM mining with different neural network

<table>
<thead>
<tr>
<th></th>
<th>CHARM</th>
<th>BPP</th>
<th>FF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time 1000</td>
<td>42</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>Memory 1000</td>
<td>33.3</td>
<td>28.8</td>
<td>18.23</td>
</tr>
<tr>
<td>Computational</td>
<td>-----</td>
<td>30%</td>
<td>11%</td>
</tr>
<tr>
<td>Complexity 1000</td>
<td>-----</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time 2000</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Memory 2000</td>
<td>22.28</td>
<td>11.784</td>
<td>7.48</td>
</tr>
<tr>
<td>Computational</td>
<td>-----</td>
<td>29%</td>
<td>12%</td>
</tr>
<tr>
<td>Complexity 2000</td>
<td>-----</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: CMSPAM mining with different neural network

<table>
<thead>
<tr>
<th></th>
<th>CHARM</th>
<th>BPP</th>
<th>FF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time 1000</td>
<td>73.2</td>
<td>36.5</td>
<td>24.7</td>
</tr>
<tr>
<td>Memory 1000</td>
<td>-----</td>
<td>32.4%</td>
<td>10%</td>
</tr>
<tr>
<td>Computational</td>
<td>-----</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complexity 1000</td>
<td>-----</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time 2000</td>
<td>1</td>
<td>0.25</td>
<td>7.5</td>
</tr>
<tr>
<td>Memory 2000</td>
<td>75.2</td>
<td>8.93</td>
<td>5.98</td>
</tr>
<tr>
<td>Computational</td>
<td>-----</td>
<td>32%</td>
<td>10.4%</td>
</tr>
<tr>
<td>Complexity 2000</td>
<td>-----</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

V. CONCLUSIONS

In this paper we study the Artificial Neural Network and how we can use ANN with the data mining concept. Computational complexity is the very important term in the any data mining field. Computational complexity defines the overall efficiency data mining algorithm. The computational complexities of neural network techniques are calculate for all the three data mining algorithms. In this paper we are successfully able to integrate the Neural Network with the three data mining techniques. Because of the neural network the output of the data mining from the all the three techniques has been improved. The result of the experiment proves our conclusion.

VI. ACKNOWLEDGEMENTS

First Author would like to acknowledge Dr. P.K.Butey for their cooperation and useful suggestion to the research work.
VII. REFERENCES


AUTHORS DETAILS

Ms. ARUNA J. CHAMATKAR is MCA from Rashtrasant Tukdoji Maharaj Nagpur University, Nagpur. Currently pursuing PhD from RTM Nagpur University under the guidance of Dr. Pradeep K. Butey. Her research area is Data Mining and Neural Network.

Dr. PRADEEP K. BUTEY is research Supervisor for Computer Science at RTM Nagpur university. He is the Head of Department (Computer Science) at Kamla Nehru Mahavidyalaya, Nagpur. His area of interest includes Fuzzy Logic, Neural Network and Data Mining.