# Implementation of Different Data mining Algorithms with Neural Network

Ms. Aruna J. Chamatkar #[1]
Research Scholar , Department Of Electronics,
& Computer Sci. RTM & Nagpur University
Nagpur , India
[1]aruna.ayush1007@gmail.com

Dr. P.K. Butey #[2]
HOD, Computer Science Department,
Kamla Nehru Mahavidyalaya
Nagpur , India
[2]buteypradeep@yahoo.co.in

***Abstract*— With the huge amount of information available online, the World Wide Web is a fertile area for data mining research. The data mining research is at the cross road of research from several research communities, such as database, information retrieval, and within AI, especially the sub-areas of machine learning and data integrity. Every E-commerce and social website in World Wide Web uses the Classification is one of the data mining problems receiving great attention recently in the database community. Neural network is not suitable for data mining directly, because how the classifications were made is not explicitly stated as symbolic rules that are suitable for verification or interpretation by humans. Different concise symbolic rules with high accuracy can be extracted from a neural network with the proposed approach. The neural network is first trained to achieve the required accuracy in data mining. 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. The different datasets of online e-commerce website filpkart and Amazon are used to train the neural network and to use in data mining. The results of all three data mining algorithm with neural network techniques then tested on the available datasets and result are compared by computational complexity of the algorithm.**

***Keywords — Artificial Neural Network, CHARM algorithm, CM SPAM Algorithm , Data Mining, K rule mining.***

## I. INTRODUCTION

It is not exaggerated to say the Web World Web is the most excited impacts to the human society in the last 10 years. It changes the ways of doing business, providing and receiving education, managing the organization etc. The most direct effect is the completed change of information collection, conveying, and exchange. The users want to have the effective search tools to find relevant information easily and precisely. The Web service providers want to find the way to predict the users' behaviors and personalize information to reduce the traffic load and design the Web site suited for the different group of users. The business analysts are expecting tools or techniques to help them satisfy their demands and/or solve the problems encountered on the Web. Therefore, Data mining becomes an active and popular research field. 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.

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. 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. DIFFERENT NEURAL NETWORK ALGORITHMS

There are many different neural network algorithms are presented by different authors in there papers. In this paper three different artificial neural network algorithms are used with data mining algorithms to improve the computational complexity of the data mining algorithm following are the three different neural network techniques used in this paper.

### A. Feedforward Neural Network

A feed forward neural network (FFNN) is the one of the simplest neural network technique, such as in Figure, consists of three

978-1-4799-6892-3/15 $31.00 © 2015 IEEE
DOI 10.1109/ICCUBEA.2015.78

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. During training this weight is adjusted. Information only travels in the forward direction through the network - there are no feedback loops.

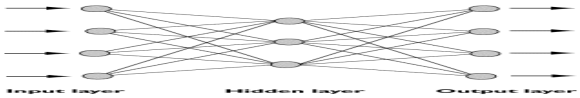

Figure 2 Flow of the feedforward neural network

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.

## B. 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)

## C. Radial Basic Function Network

The idea of Radial Basis Function (RBF) Networks derives from the theory of function approximation. We have already seen how Multi-Layer Perceptron (MLP) networks with a hidden layer of sigmoidal units can learn to approximate functions. RBF Networks take a slightly different approach. Their main features are:

1. They are two-layer feed-forward networks.
2. The hidden nodes implement a set of radial basis functions (e.g. Gaussian functions).
3. The output nodes implement linear summation functions as in an MLP.
4. The network training is divided into two stages: first the weights from the input to hidden layer are determined, and then the weights from the hidden to output layer.
5. The training/learning is very fast.
6. The networks are very good at interpolation.

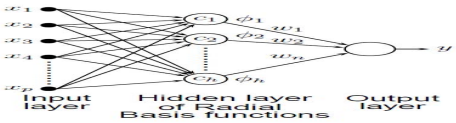


Figure 3 Radial Basis Function Network

The radial basis functions in the hidden layer produce a significant non-zero response only when the input falls within a small localized region of the input space. Each hidden unit has its own receptive field in input space. An input vector xi which lies in the receptive field for center $c_j$ , would activate $c_j$ and by proper choice of weights the target output is obtained. The output is given as:

$$y = \sum_{j} w_j \varphi_j \quad , \quad \varphi_j = \varphi(\|x - c_j\|)$$

$w_j$ : Weight of $j^{th}$ center, $\varphi$: some radial function.

# 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.

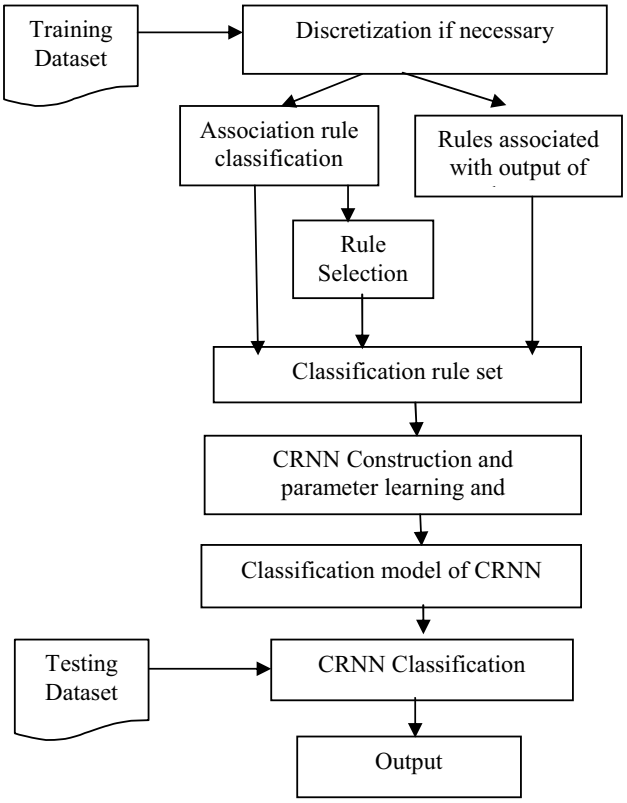


Figure 4 Association rule mining with neural network

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.

## B. 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.

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) 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.

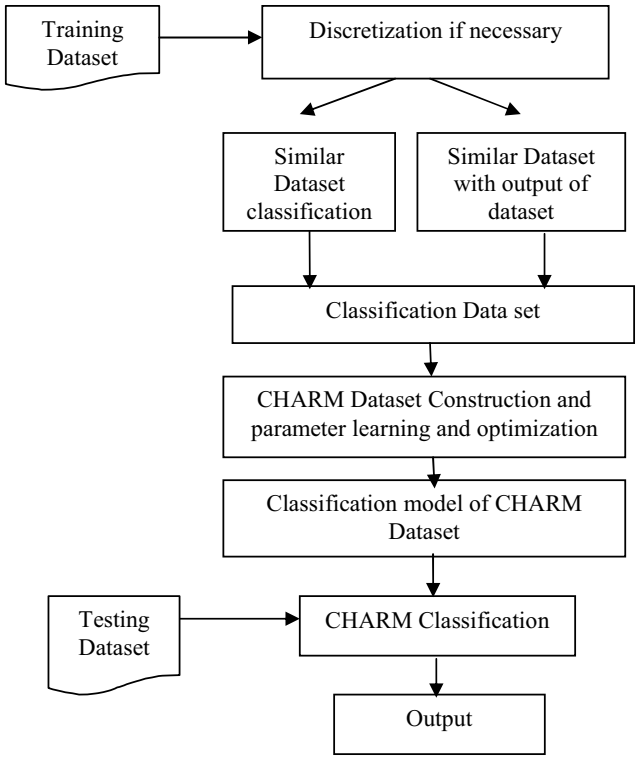


Figure 5 CHARM algorithms with neural network

The integration of the CHARM algorithm with the neural network is shown in figure 5.

## C. 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.

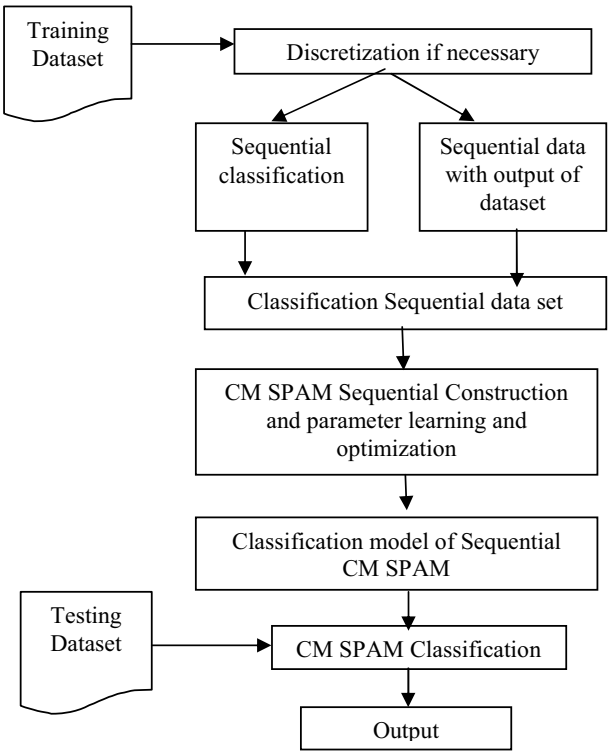


Figure 6. CM SPAM algorithms with neural network

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.

## V.EXPERIMENTAL RESULT

Once In our research we have studied various algorithms, and found out the best ones in the field of data-mining. These algorithms are namely, CHARM Bit Set for Frequent Pattern Mining, Top K Rules for Association Rule Mining and CMSPAM for Sequential Pattern Mining. We collected demo datasets of online retailing e-commerce websites Flipkart and Amazon. Each dataset contains 1000, 2000, 5000 and 10000 entries, and is available for Frequent Pattern Mining, Association Rule Mining and Sequential Pattern Mining. We applied these datasets to each of the algorithms and found out the results in terms of memory and time complexity.

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, Feed Forward with Back Propagation and Radial Basis Networks. 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, 2000, 5000 and 10000 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

| | **Top K Rules** | **BPP** | **CFF** | **RBF** |
|---|---|---|---|---|
| **Time 1000** | 24 | 31 | 7 | 11 |
| **Memory 1000** | 8.28 | 5.01 | 2.83 | 2.98 |
| **Computational Complexity 1000** | ----- | 32% | 13% | 16% |
| **Time 2000** | 171 | 119 | 58 | 76 |
| **Memory 2000** | 15.489 | 7.62 | 4.72 | 3.03 |
| **Computational Complexity 2000** | ----- | 31% | 13.2% | 17% |
| **Time 5000** | 133 | 54 | 51 | 67 |
| **Memory 5000** | 10.48 | 6.47 | 5.11 | 2.69 |
| **Computational Complexity 5000** | ----- | 31.7% | 12.9% | 16.5% |
| **Time 10000** | 235 | 125 | 65 | 89 |
| **Memory 10000** | 16.92 | 7.63 | 6.23 | 5.14 |
| **Computational Complexity 5000** | ----- | 33% | 13% | 16% |

Table 2:CHARM Mining with different neural network techniques

| | CHARM | BPP | CFF | RBF |
|---|---|---|---|---|
| **Time 1000** | 42 | 13 | 10 | 19 |
| **Memory 1000** | 33.3 | 28.8 | 18.23 | 14.1 |
| **Computational Complexity 1000** | ----- | 30% | 11% | 14% |
| **Time 2000** | 1 | 1 | 1 | 2 |
| **Memory 2000** | 22.28 | 11.184 | 7.48 | 5.63 |
| **Computational Complexity 2000** | ----- | 30.5% | 11.6% | 15% |
| **Time 5000** | 7 | 2 | 1 | 3 |
| **Memory 5000** | 24.11 | 19.65 | 13.129 | 9.864 |
| **Computational Complexity 5000** | ----- | 29% | 12.2% | 16% |
| **Time 10000** | 11 | 3 | 3 | 6 |
| **Memory 10000** | 56.1 | 28.6 | 23.7 | 17.86 |
| **Computational Complexity 10000** | ----- | 31% | 13% | 16.2% |

Table 3: CMSPAM Mining with different neural network techniques

| | CMSPAM | BPP | CFF | RBF |
|---|---|---|---|---|
| **Time 1000** | 2 | 0.5 | 1.5 | 0.5 |
| **Memory 1000** | 73.2 | 36.5 | 24.7 | 18.38 |
| **Computational Complexity 1000** | ----- | 32% | 10% | 18% |
| **Time 2000** | 1 | 0.25 | 7.5 | 0.5 |
| **Memory 2000** | 75.2 | 8.93 | 5.98 | 4.508 |
| **Computational Complexity 2000** | ----- | 31.6% | 12.3% | 17.4% |
| **Time 5000** | 2 | 0.5 | 0.5 | 0.5 |
| **Memory 5000** | 19.121 | 9.6 | 6.44 | 4.855 |
| **Computational Complexity 5000** | ----- | 31% | 12% | 17.1% |
| **Time 10000** | 1 | 0.75 | 0.5 | 0.25 |
| **Memory 10000** | 20.3 | 10.3 | 6.89 | 5.193 |
| **Computational Complexity 10000** | ----- | 31.2% | 11.7% | 17% |

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. As shown in table 1, table 2 and table 3, the computational complexity for the Top K Rule mining with Back Propagation is 32 %, with Radial Basis Function Networks is 16% and with Feed Forward is 13%. Computational complexity for CHARM Mining with Back Propagation is 30%, with Radial Basis Function Networks is 14% and with Feed Forward is 11%. Computational complexity for CHSPAM Mining with Back Propagation is 32%, with Radial Basis Function Networks is 18% and with Feed Forward is 10%. So the computational complexity of the Feed Forward technique is less than the other technique of neural network for the data mining algorithms.

## IV. CONCLUSION

In this paper we study the Artificial Neural Network and how we can use ANN with the data mining concept. In this paper we are successfully able to integrate the Neural Network with the three data mining techniques. From the above result its clear that time and memory complexity does not entirely depend on the length of the input dataset. For frequent pattern mining Radial Basis and Cascaded Feed-forward Networks are the best. Association Rule Mining follows the same trend of output as Frequent Pattern mining; Radial Basis and Cascaded Feed-forward Networks are the best. Sequential pattern mining is best done by Radial Basis Networks.

ACKNOLEDGMENT

First Author would like to acknowledge Dr. P.K.Butey for their cooperation and useful suggestion to the research work and Dr. A. K. Shende Principle of Kamla Nehru Mahavidyalaya, Nagpur.

## REFERENCES

[1] Neelamadhab Padhy, Dr. Pragnyaban Mishra, Rasmita Panigrahi "The Survey of Data Mining Applications And Feature Scope" at International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012.

[2] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.

[3] Larose, D. T., "Discovering Knowledge in Data: An Introduction to Data Mining", ISBN 0-471-66657-2, ohn Wiley & Sons, Inc, 2005.[4] Dunham, M. H., Sridhar S., "Data Mining: Introductory and Advanced Topics", Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006

[5] M. J. Zaki and K. Gouda. Fast vertical mining using Diffsets. Technical Report 01-1, Computer Science Dept., Rensselaer Polytechnic Institute, March 2001.

[6] Mabroukeh, N.R., Ezeife, C.I.: A taxonomy of sequential pattern mining algorithms. ACM Computing Surveys 43(1), 1–41 (2010).

[7] Agrawal, R., Ramakrishnan, S.: Mining sequential patterns. In: Proc. 11th Intern. Conf. Data Engineering, pp. 3–14. IEEE (1995)

[8] Bradley, I., Introduction to Neural Networks, Multinet Systems Pty Ltd 1997.

**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 the Department (Computer Science) at Kamla Nehru Mahavidyalaya, Nagpur . His area of interest includes Fuzzy Logic ,Neural Network and Data Mining