Combining Feature Selection and Neural Networks for Solving Classification Problems

Jigyasa Kohli (jk423), Mahesh Muraleedharan Nair (mmn9), Venkata Jayasimha Hari Chidiri (vc279)
Group 3, MGMT 635851
Information Systems Department,
New Jersey Institute of Technology, Newark.

Abstract

In this paper, an approach to solving classification problems by using neural networks system is elaborated. Field of Information theory is used to select a set of important attributes that can be used to classify tuples. A neural network is trained using these attributes; the neural network is then used to classify tuples. Data mining topics will be discussed and a well known dataset of German Credit card system will be used to create a neural network and churn this system's performance.

1. INTRODUCTION

Data mining is defined as the process of discovering patterns in data.The process must be automatic or semiautomatic ;mostly the processes are semiautomatic.The patterns discovered must be meaningful , when they have some meaning they provide some advantage which is mostly seen as an economic advantage.So, Data mining is all about solving problems by analyzing data already present in databases.In today's highly competitive,customer centered data is the raw material that fuels business growth but for this data should be properly mined.The data is constantly present in big quantity which is refered as Big Data.

There has been stunning progress in data mining and machine learning.The synthesis of statistics, machine learning , information theory , and computing has created a solid science , with a firm mathematical base, and with very powerful tools.

This paper is an exploration of neural networks for pattern recognition in German Dataset.This is followed by specifically training of data and by repeated exposure to data , so that the network can be used to make pragmatic decisions.When they are trained with samples of input-output data they can make predictions,classifications on the basis of performance,confusion matrix, error histogram.The above have been discussed in the upcoming topics.

2. DATA MINING TASKS

Data mining deals with the kind of patterns that can be mined. On the basis of the kind of data to be mined, there are two categories of functions involved in Data Mining −

- Descriptive
- Classification and Prediction

But this paper mainly focuses on the classification and prediction which is based on the analysis of set of training presented in the form of neural networks.

Descriptive Function

The descriptive function in Data Mining deals with the general attributes and character of data in the databases. The list of descriptive functions comprises of :

- Class/Concept Description
- Mining of Frequent Patterns
- Mining of Associations
- Mining of Correlations
- Mining of Clusters

Classification and Prediction Function

Classification function in Data Mining is the process of finding a prototype that describes the data classes or concepts. Suppose, if there is a class of objects whose label is unknown then we can use this model to predict the class.This model is based on the analysis of sets of training data. The various forms of this model are : Classification (IF-THEN) Rules,Decision Trees,Mathematical Formulae,Neural Networks.

The various functions which are involved in these processes are −

- **Classification** − It predicts the class of objects whose class label is unknown. Its objective is to find a derived model that describes and distinguishes data classes or concepts. The Derived Model is based on the analysis set of training data i.e. the data object whose class label is well known.
- **Prediction** − It is used to predict missing or unavailable numerical data values rather than class labels. Regression Analysis is generally used for prediction. Prediction can also be used for identification of distribution trends based on available data.
- **Outlier Analysis** − Outliers may be defined as the data objects that do not comply with the general behavior or model of the data available.
- **Evolution Analysis** − Evolution analysis refers to the description and model regularities or trends for objects whose behavior changes over time.(https://www.tutorialspoint.com/data_mining/dm_tasks.htm, October 8,2016).

2.1 Applications

Data mining has been used extensively in the area of market and sale.These are the domains in which companies have myriad volumes of precisely recorded data which is extremely valuable, this is a part of Big data.In these applications,predictions are the most important part.Banks were using the Data Mining concept since a long time because of their success in the use of machine learning for credit assignment.Data Mining is now being used to group customers like in acluster of profitabe,reliable customers and many more clusters by detecting change in individual banking patterns.

3.TECHNIQUES AND APPROACHES

a. Decision Tree

Decision tree learning is a method that uses inductive inference to approximate a target function, which will produce discrete values. It is widely used, robust to noisy data, and considered a practical method for learning disjunctive expressions.

"In general, decision trees represent a disjunction of conjunctions of constraints on the attribute-values of instances. Each path from the tree root to a leaf corresponds to a conjunction of attribute tests, and the tree itself to a disjunction of these conjunctions" (Mitchell, 1997, p.53). Until each leaf node is populated by as homogeneous a sample set as possible:

General Form

- Select a leaf node with an inhomogeneous sample set.
- Replace that leaf node by a test node that divides the inhomogeneous sample set into minimally inhomogeneous subsets, according to an entropy calculation.

Specific Form

- Examine the attributes to add at the next level of the tree using an entropy calculation.
- Choose the attribute that minimizes the entropy.

The central focus of the ID3 algorithm is selecting which attribute to test at each node in the tree.

Procedure:

1. See how the attribute distributes the instances.
2. Minimize the average entropy.
    - Calculate the average entropy of each test attribute and choose the one with the lowest degree of entropy.

- ID3 searches a *completely* expressive hypothesis space (ie. one capable of expressing any finite discrete-valued function), and thus avoids the difficulties associated with restricted hypothesis spaces.
- ID3 searches *incompletely* through this space, from simple to complex hypotheses, until its termination condition is met (eg. until it finds a hypothesis consistent with the data).
- ID3's inductive bias is based on the ordering of hypotheses by its search strategy (ie. follows from its search strategy).
- ID3's *hypothesis space* introduces no additional bias.

b. Clustering

Cluster analysis is the process of grouping objects into subsets that have meaning in the context of a particular problem. The objects are thereby organized into an efficient representation that characterizes the population being sampled. Unlike classification, clustering does not rely on predefined classes. Clustering is referred to as an ***unsupervised learning method*** because no information is provided about the "right answer" for any of the objects. It can uncover previously undetected relationships in a complex data set. Many applications for cluster analysis exist.
For example, in a business application, cluster analysis can be used to discover and characterize customer groups for marketing purposes.
Two types of clustering algorithms are ***nonhierarchical*** and ***hierarchical***. In nonhierarchical clustering, such as the **k-*means*** algorithm, the relationship between clusters is undetermined.

Hierarchical clustering repeatedly links pairs of clusters until every data object is included in the hierarchy. With both of these approaches, an important issue is how to determine the similarity between two objects, so that clusters can be formed from objects with a high similarity to each other. Commonly, ***distance functions***, such as the ***Manhattan*** and ***Euclidian*** distance functions, are used to determine similarity. A distance function yields a higher value for pairs of objects that are less similar to one another. Sometimes a ***similarity function*** is used instead, which yields higher values for pairs that are more similar.

***k*-means Algorithm:**

1. Select *k* clusters arbitrarily.
2. Initialize cluster centers with those *k* clusters.
3. Do loop.

a) Partition by assigning or reassigning all data objects to their closest cluster center.
b) Compute new cluster centers as mean value of the objects in each cluster.
Until no change in cluster center calculation.

c. Neural Networks

Neural networks have been successfully applied in a wide range of supervised and unsupervised learning applications. Neural-network methods are not commonly used for data-mining tasks, however, because they often produce incomprehensible models and require long training times. In this article, we describe neural-network learning algorithms that are able to produce comprehensible models, and that do not require excessive training times.

Anatomy Of Neural Network

1.Neural Network map a set of input nodes to a set of output nodes.
2. Number of Input Nodes and Output Nodes is variable.
3.The Network itself is composed of arbitrary number of nodes with arbitrary topology.

Neural networks are used in a wide variety of applications. They have been used in all facets of business from detecting the fraudulent use of credit cards and credit risk prediction to increasing the hit rate of targeted mailings.

The neural network used in paper is used to extract features by requiring the network to learn to recreate the input data at the output nodes by using different number of hidden nodes.
A network can be trained to map input values to corresponding output values by providing a training set. The network is repeatedly tested and modified to produce the correct output.
The generation of output by a neural network is accomplished via firing values from nodes. An input is passed to the input layer which in turn can activate the internal layers, which in turn activates the output layer, finally resulting in an output.

Neural Network follows the topology of Back Propagated Network in which inputs are put through a 'hidden layer 'before the Output Layer.All the nodes are connected between the layers.The supervised training of Back Propagated Network includes :
1.Desired output of training inputs.
2.Error= Difference Between Desired and actual Output.
3. Change Weight for more accuracy.
4. Changing the output layer by propagating back to previous layer.
5. Hidden Layers and Number of neurons in the network.

In the data we had 1,000 observations. We trained, tested and validated a neural network with the first 980 observations as training inputs and rest 20 observations as testing inputs and by changing number of neurons in the hidden layer. 20 observations were kept as holdout outputs which was compared with the output we got after training the inputs.

Information Theory

Information theory measure information in bits.

Information gain=(Entropy of distribution before the split)–(entropy of distribution after it)

Information gain is the amount of information that's gained by knowing the value of the attribute, which is the entropy of the distribution before the split minus the entropy of the distribution after it. The largest information gain is equivalent to the smallest entropy.

The entropy (very common in Information Theory) characterizes the (im)purity of an arbitrary collection of examples.

Entropy Calculations

If we have a set with k different values in it, we can calculate the entropy as follows:

Entropy(p1,p2,…,pn)=−p1log(p1)−p2log(p2)− −pnlog(pn)

If all instances in a group were known to be all in the same class, then the information value of being told the class of a particular instance is Zero

If instances are evenly split between classes, then the information value of being told the class of a particular instance is Maximized.

Information theory measures the value of information using "***entropy***", which is measured in "***bits***".

For example, if we assume in our German Data and take a single attribute Credit History ,let's take the dataset 30 in Credit History then let's find its Entropy and Information Gain :

Child Entropy= *-25/40 [log(25/40)/log 2] - 15/40 [log(15/40)/log2]*

*=0.68 *0.625 + 0.375 * 1.4 (Rounded Figure)*

=0.9544

Now, let's find the weight of number of Customer who has credit history as 30 as follows :

Weight of each value = *-P(Customers with credit history 30) * [log P(Customers with credit history 30)/log 2]*

*= -0.04 * [log (0.04)/log 2)*

*=-0.04 * -4.6*

= 0.18 (by rounding off)

Weighted Entropy = *P(Customers with credit history 30) * Child Entropy*

*= 0.04 * 0.9544*

=0.0381(Rounded Figure)

After all the calculations

Information Gain = *Entropy of all the data - Entropy Of Child Expected Data .*

Here, We found Entropy of all the data was 0.8812 and Conditional Entropy of Credit History set is 0.837

So,Information gain for Credit History =0.8812-0.837 = 0.436.

4. APPROACH

We followed the approach of **Cross-Industry Standard Process** for Data Mining (CRISP-DM). For German Credit Dataset.

4.A) BUSINESS UNDERSTANDING

- **Determining the business objective** – Applicant credit risk analysis
- **Assess situation** – Good credit and Bad credit applicant
- **Determining data mining goals**- Analyzing loan applicant attributes like status, duration, credit history ,credit amount, savings, housing, foreign worker.
- **Project planning** – Importing Data from Excel Sheet to MATLAB and then dividing that data into matrix; Input Data as Training Input(980*20) and Testing Input Data (20*20) and Output as training Output (980 *1) and Testing Output Data (20*1), then select Pattern recognition.Then,after that running simple code, then interpreting our charts.This is followed by taking 20 criteria and then eliminating 13 and carrying out with 7 criterias.

4.B) DATA UNDERSTANDING:

- **Collect initial data**- German credit data source.
- **Describe data**- 1000 applicants with 20 attributes like status, duration, credit history, credit amount, savings, housing, foreign worker.In addition to this a column of data to describe whether an applicant is a good or bad customer.
- **Explore data**- Selected nntool for the analysis and used npr tool for pattern recognition . Visualization of the data was done by Graph Analysis like in the form of Histogram Plots and Transforming the 20 attributes to 7 attributes which have a significant impact on accuracy.
- **Verify data quality** - Data was recorded correctly in MATLAB and all the relevant data was recorded. Validating whether all the uploaded file fit an expected pattern. Any abnormalities in the data has been cleaned out. The correct codes are used as per the rest of the dataset.

4.C) DATA PREPARATION:

- **Select data-** 980 rows with 20 attributes
- **Cleaning data-** Any abnormalities in the data has been cleaned out. Out of 20 attributes describing each pattern in the german dataset, 13 attributes were eliminated.
- **Construct data-** 7 attributes status, duration, credit history, credit amount, savings, housing and foreign worker were selected and data is trained again .
- **Integrate data-** combining using merging tables.

4.D) MODELLING:

- **Modeling technique-** Neural networks
- **Test design-** Training Datasets and Test Datasets .
- **Built model-** Training Input Sets - 980 * 20, Training Output Sets - 980*1, Testing Output set - 20 * 1, evaluation set - 20 * 20, score set - 20 * 1
- **Assess model-** Confusion Matrix

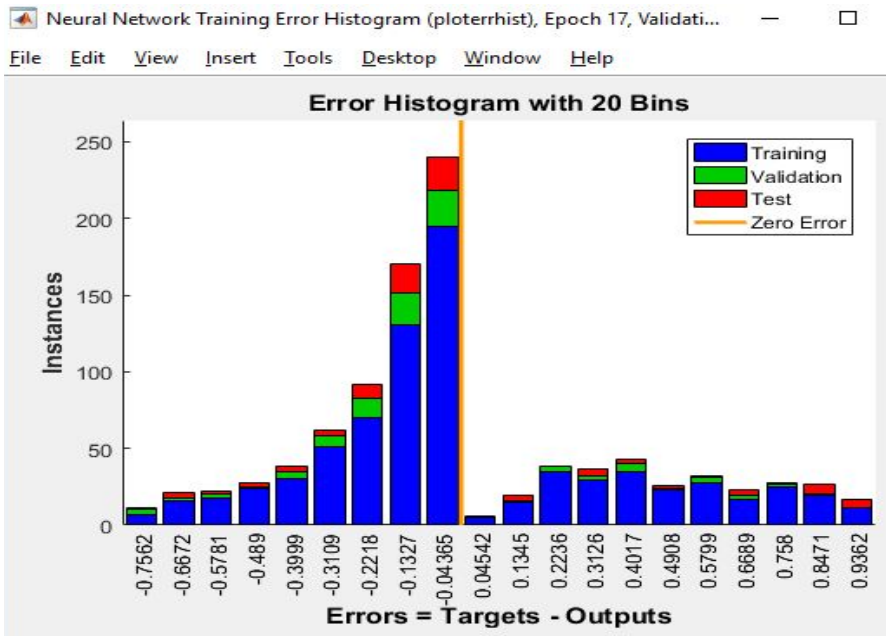
4.E) TESTING AND EVALUATION:

Two evaluations were noted to return best results under different parameters;

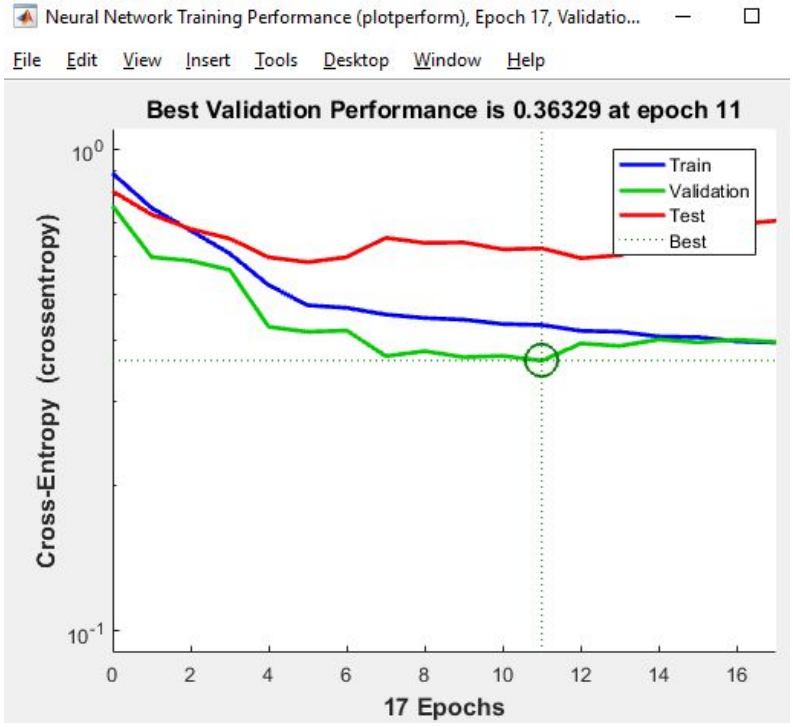
a. **16 of the holdout samples were right when hidden layer is 10** AND
b. **17 of the holdout samples were right when hidden layer is 20**

This result (**b.**) will be focused on, as it is the best experimental result under the several tests done.

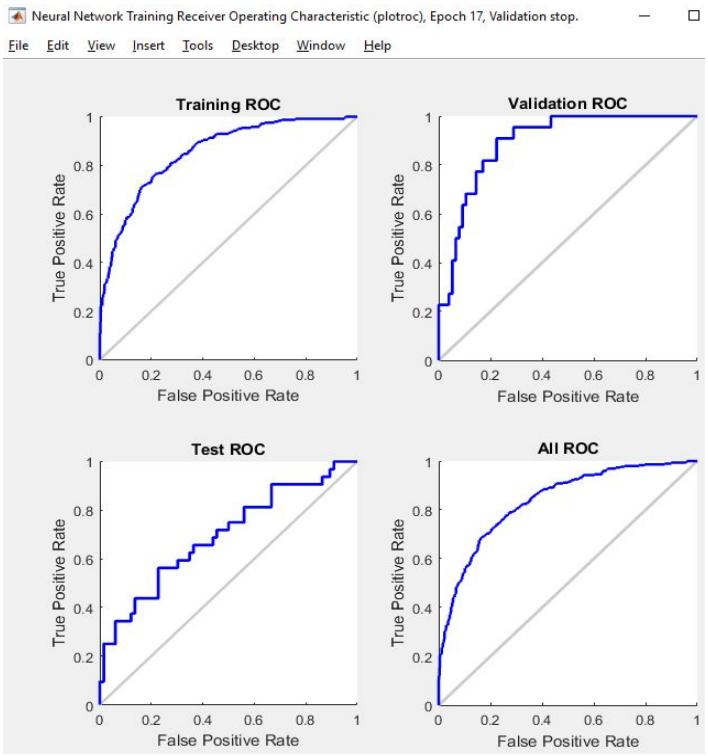
Plot 1. *Confusion plot*



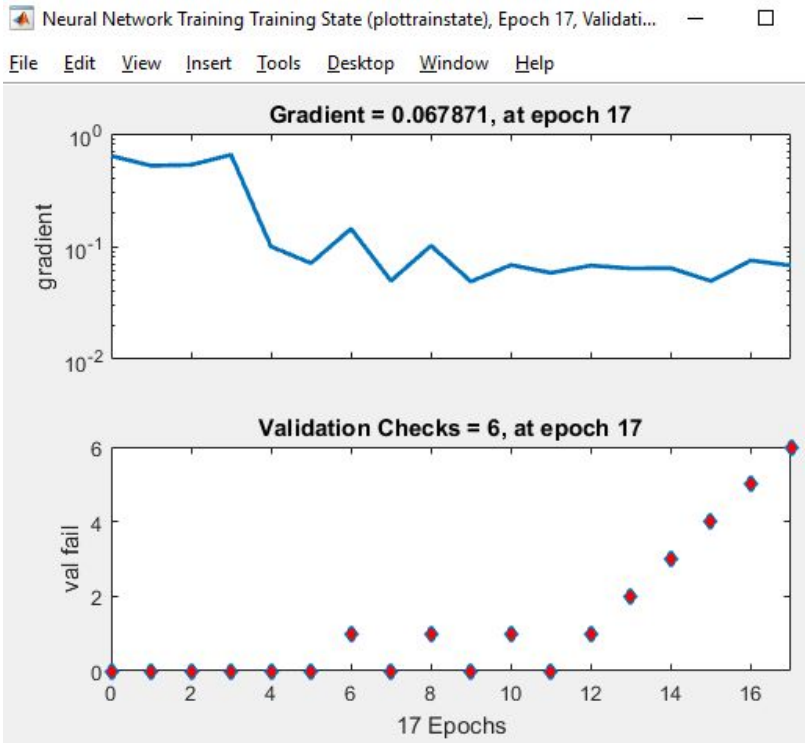
Plot 2. *Error Histogram Plot*



Plot 3. *Performance plot*



Plot 4. *Receiver Operating Characteristic plot*



Plot 5. *Training State plot*

4.F) DEPLOYMENT:

The results are very inconsistent and Neural Network is not a technique that can be considered as the go to to system in a credit card company for good and bad credit. It can be used though to do regressive prediction along with another analytical technique to cross check True Positives. Maybe training with several thousand datasets could improve accuracy of the predictions.

5. RESULTS AND DESCRIPTION OF RESULTS

The experiments were conducted using the German Dataset with the help of neural network training.The Dataset consisted of the information on 1000 applicants and they are having a set of 20 different attributes.These 20 attributes were divided as continuous(3) and discrete attributes(17).

The last column on the classification is whether a customer credit rating is actually "good" or "bad" .Thus the prediction is to be done to classify each pattern good or bad.
In the data set, there are a total of 700 cases of good applicants and 300 cases of bad applicants. Furthermore, the data set is divided into a training set and a test set. Training Input(980*20) and Testing Input Data (20*20) and Output as Training Output (980 *1) and Testing Output Data (20*1).

We improved the accuracy by changing various parameters, such as number of neurons, and the number of layers. Using the neural network and its algorithm from the 20 attributes describing each pattern in the german dataset,13 were being eliminated and 7 were selected for further improvement of accuracy. In Table 1, we list the information gains , normalised gains and their averages for the german credit problem. The seven attributes which were selected are: status, duration, credit history, credit amount, savings, housing and foreign worker.

No.	Attribute	$\mathscr{G}$	$\mathscr{G}'$
1.	Status	0.094738842	0.052573017
2.	Duration	0.063354593	0.016828623
3.	Credit history	0.043617799	0.025479578
4.	Purpose	0.02489354	0.009335039
5.	Credit amount	0.019340301	0.009670151
6.	Savings	0.028114675	0.016658196

7.	Present employment	1.31023E-02	0.010787589
8.	Installment	3.97213E-03	0.002195826
9.	Personal status and sex	6.81055E-03	0.004445228
10.	debtors / guarantors	4.79702E-03	0.008908709
11.	Present residence	0.000542501	0.000294505
12.	Property	0.016985186	0.008720275
13.	Ages	0.045585571	0.008650556
14.	Other Installation Plan	0.00887507	0.010506605
15.	Housing	0.012753186	0.011196712
16.	No. of existing credits at this bank	0.001978331	0.00174364
17.	Job	0.001337357	0.000946195
18.	Number of people being liable to provide maintenance	6.56762E-06	1.05553E-05
19.	Telephone	9.63660E-04	0.000990154
20.	Foreign worker	5.82299E-03	0.025498723
Average		**0.019879608**	**0.011271994**

Table 1. 20 Attribute gains of the german credit data set.

After 7 attributes were selected for further improvement of accuracy we calculate the normalized information gain

Difference of Normalized Information Gain=(7 attribute Normalized Information Gain) - (20 attribute Normalized Information Gain)
=0.0225578571-0.011271994
=0.0112858631

Accuracy=number of tuples correctly classified / Total number of tuples

Due to the presence of noise, it is not possible to achieve 100% accuracy on training data for the german credit problem. But on training the network again and again and simulating the network the accuracy was improved which was found out to be **85%**.There is always complexities with real world data and other reason can be there may be lack of balance in data due to repetitive data and inconsistency of data,missing data that is the reason for the less accuracy in the result.

Attribute	Input Number
status	$\mathscr{I}_1$ - $\mathscr{I}_4$
duration	$\mathscr{I}_5$ - $\mathscr{I}_{38}$
Credit history	$\mathscr{I}_{39}$ - $\mathscr{I}_{44}$
Credit amount	$\mathscr{I}_{45}$ - $\mathscr{I}_{49}$
savings	$\mathscr{I}_{50}$ - $\mathscr{I}_{54}$
housing	$\mathscr{I}_{55}$ - $\mathscr{I}_{57}$
Foreign worker	$\mathscr{I}_{58}$ - $\mathscr{I}_{60}$

Table 2. Binarization of the attribute values

6.Conclusions

The experiment had iterate the training of data sets several times by changing the hidden layer, the output was most accurate at hidden layer 20, with yield of 17/20 and a close second was with hidden layer 10, with yield of 16/20. Consideration of hidden layer 20 makes sense because of its 85% accuracy. Manipulating the hidden layer could yield more accurate results.However over a period of multiple iteration, we could find some inconsistencies in the accuracy probably because of mismatched data in the training set not been trained into the pattern recognized training set.

The attributes were reduced to 7 from 20 but the predictions from the trainings are still inconsistent.

References

[1] https://www.tutorialspoint.com/data_mining/dm_tasks.htm
[2] Alex Berson, Stephen Smith, and Kurt Thearling ,**Building Data Mining Applications for CRM**
[3] Credit scoring in banks and financial institutions via data mining techniques: By Seyed Mahdi sadatrasoul ,Mohammadreza gholamian,Mohammad Siami,Zeynab Hajimohammadi
[4] Kuhn, M., and Johnson, K. (2013). Applied Predictive Modeling. Springer.
[5]S. M. Kamruzzaman and A. M. Jehad Sarkar .A New Data Mining Scheme Using Artificial Neural Networks