Analysis of Banking Data Using Machine Learning

Priyanka S. Patil
Department of Computer Science and Engineering
Rajarambapu Institute of Technology, Islampur, India
Email: priyapatil3131@gmail.com

Nagaraj V. Dharwadkar
Department of Computer Science and Engineering
Rajarambapu Institute of Technology, Islampur, India
Email: nagaraj.dharwadkar@ritindia.edu

***Abstract*- In today's world large amount of data is generated in every field and banking industry is one of them. This data contains valuable information. Hence, it is very important to store, process, manage and analyze this data to extract knowledge from it. It helps to increase business profit. Banking industry plays very important role in economy of country. Customers are the main asset of the bank. Hence it is necessary to focus on problems faced by the banks. Here, we are working on customer retention and fraud detection. In this work, supervised artificial neural network algorithm is implemented for classification purpose.**

Keywords- Banking industry, Customer Retention, Fraud Detection, Machine Learning, Artificial Neural Network

I. INTRODUCTION

The Banking industry generates a massive volume of data every day. It contains customer account information, transaction information, all financial data etc. Data analytics can be used to analyze large volume data to extract meaningful information from it [7]. It helps to uncover hidden information, hidden patterns and to discover knowledge from the large volume data [12]. Banks are facing various challenges like customer retention, fraud detection, risk management [2] and customer segmentation [1]. It needs to focus on these challenges to increase business profit. Customer retention is effective method for the growth of the banks. In the banking sector churn and fraud becomes major problem today [8]. So it is important to identify customer's behavior and retain them. To retain customers first it is necessary to identify which customers are active and inactive. Machine learning helps to handle large data in the most intelligent fashion by developing algorithms to generate insights from it [14]. Here bank customer data is used. In this work we are using supervised artificial neural network to perform classification.

II. RELATED WORK

A literature review gives many results on analysis of banking and financial data which was carried out by different methods, techniques. Many researchers have developed and implemented various analysis and prediction models using different data mining techniques.

Yong Shic et al. discussed customer churn prediction in commercial bank. They used Support Vector Machine algorithm for classification purpose. To improve the performance of the SVM model random sampling method is used and F-measure is selected for evaluating predictive power in this paper. They also developed Logistic regression model and made comparisons between developed models. The results clears that the SVM model random with sampling method works better [5].

Iain Brown et al. compared different techniques used for analysis of credit scoring datasets. They have compared results of classification techniques like neural network, logistic regression, gradient boosting, random forests and least square support vector machines. They have checked the performance of techniques with increasing class imbalance in datasets using under sampling methods. They concluded that Random forest and gradient boosting is works better as compared to other techniques in case of large data imbalance [4].

A.B. Adeyemo et al. predict churners by using data mining techniques. They have used real-life customer records given by Nigerian bank. They have first cleaned and preprocessed raw data and then analyzed using WEKA tool. For clustering simple K-Means was used and rule based algorithm, JRip was used for

the rule generation purpose. The obtained result shows that the implemented methods can determine patterns in customer behaviors and helps to banks in identification of churners [9].

Dr. U. Devi Prasad et al. studied purchasing patterns of bank customers in Indian industries. They developed model to convert raw customer data into useful data. They have used data mining techniques to design model. They have experimented with classification methods which are CART, and C 5.0. The CART is predicting churn class successfully than C 5.0. The C 5.0 is predicting active class successfully than CART [11].

Cheng-Tao Chu et al. implemented Multilayer Perceptron (MLP) neural networks with back-propagation learning for churn prediction in a telecommunication company. To build classification model they have used different topologies of MLP. For this they used real data of customers in a major Jordanian telecom company. They evaluated and compared two methods which are typical change on error and the ANN weights based method [15].

There are various machine learning algorithms used for classification or clustering in literature by many authors. Table 1 describes some algorithms like Decision Tree, SVM, Neural Network, and K-means.

Table 1: Different Machine Learning Algorithms used for classification and clustering

Machine Learning Algorithms	**Description**
Decision Tree (DT)	It generates rules or patterns which are easy to interpret. These rules are in the form of if-then-else expressions.
K-means	It divides data into clusters on the basis of centroid. Elements in same cluster are close to centroid of that cluster.
Naive Bayes (NB)	It is used for making predictions. It uses Bayes' Theorem. It derives the probability of a prediction according to events present in the data [14].
Support Vector Machine (SVM)	It uses various kernel functions to process different data. Kernels functions are linear and nonlinear [14].
Neural Network	It is distributed processing systems and it has ability to learn complex patterns presents in the data. Accuracy rate of this algorithm is also better.

III. PROPOSED METHOD

A. System Architecture

The system architecture includes various phases like data collection, data pre-processing, making training and testing dataset, implementing ANN algorithm and result analysis. Proposed system architecture is shown in figure 1.

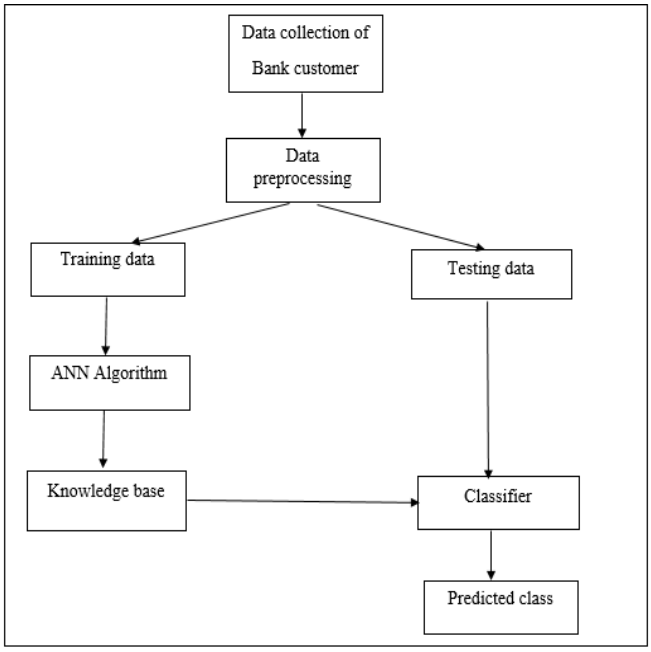

Figure 1: System Architecture

First bank customer data is prepared for processing. Input dataset contains both types of values categorical and numerical. To implement artificial neural network algorithm first all data is converted into numerical values and then used as input to algorithm for processing. Hence first data is preprocessed for processing. The data is divided into two parts training data and testing data to check performance of the model.

Here we have used Artificial Neural Network as a machine learning algorithm for classification and prediction. ANN can process multiple inputs efficiently and also it handles large, complex data easily [3]. Hence this algorithm is used. First algorithm is applied on training data and

prepared model. Then model is applied on testing data which is not used for training.

B. Working of Artificial Neural Network Algorithm

The structure of ANN consists of three basic layers such as input, hidden and output layer. Following figure 2 shows structure of neural network.

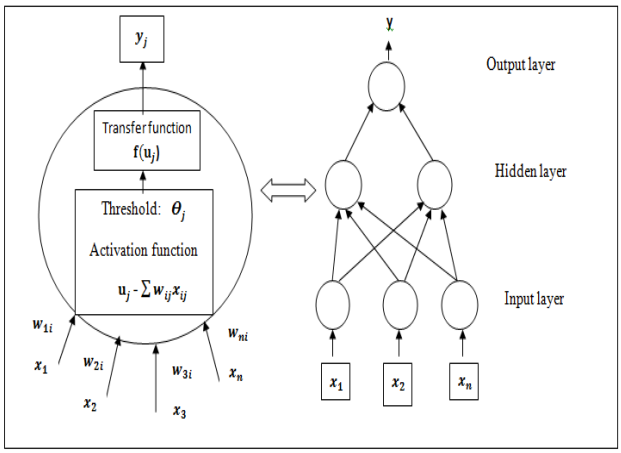

Figure 2: A Neuron and Artificial Neural Network

a. Components of Neural network

1. Weighting Factors: In ANN neuron receives multiple inputs simultaneously. To process element's summation function weight is assigned to every input. Weights are adaptive coefficients which help to decide the intensity of the input given to neural network.

2. Summation Function: The inputs and its respective weights can be represented as $(i_1,i_2,\ldots i_n)$ and $(w_1,w_2,\ldots w_n)$. The total input is the dot product of these two vectors. The result is $(i_1 * w_1) + (i_2 * w_2) + \ldots\ldots + (i_n * w_n)$. The summation function can be complex. The input and weights can be combining in various ways before passing to the transfer function

3. Activation/Transfer function: algorithmic process it means transfer function is applied on output of the summation function and it transformed to a working output.
There are many possible activation functions

i. Step function: The output of this function is binary. It depends on whether the input meets a specified threshold, Θ. If the activation meets the threshold then output is set to one.

$$f(n) = \begin{cases} 0, n < \Theta \\ 1, n \geq \Theta \end{cases} \quad \ldots\ldots\ldots (1)$$

ii. Log-sigmoid function: This function is the most popular. The sigmoid function is a S-shaped graph. This function takes the input and provides the output into the range 0 to 1, according to the expression:

$$f(n) = \frac{1}{1+e^{-n}} \quad \ldots\ldots\ldots (2)$$

iii. Hyperbolic tangent sigmoid: The tan-sigmoid transfer function produce output ranges between –1 and 1, according to the following equation:

$$f(n) = \frac{e^{n}-e^{-n}}{e^{n}+e^{-n}} \quad \ldots\ldots\ldots (3)$$

4. Scaling and Limiting: If transfer function is completes, the result can pass through additional processes like scale and limit.

5. Output Function: Each input element has one output associated with it. Normally, the output is equivalent to the result of transfer functions.

6. Error Function: In the process of learning neural networks, difference between the current output and the actual output is evaluated as an error and then it transformed by error function. This error is propagated backwards to a previous layer to update weights.

7. Learning Function: It is used to change the weights on the inputs of each processing element. The learning rate decides how adjustment is required to make network better.

C. ANN Back Propagation Algorithm (BP)

For feed forward networks the most common learning algorithm is the back-propagation (BP) algorithm. In unsupervised training models, training set contains only input vectors. There is no output vector is associated with it. The output is decides by the network. In supervised neural networks, both input and its corresponding output is provided with it and networks are rewarded for correct classification.

This algorithm consists of two phases which are propagation phase and weight update phase. In first phase, activations are forwarded from input to output layer with adding weights to neurons and computes activation function. Then model calculates error which is difference between actual and the predicted value. This error is propagated backward to modify weights. In second phase ANN weights are adjusted to reach minimization criteria [6].

Algorithm: Artificial Neural Network

Input: Dataset with input and corresponding output
Output: Predicted class

1. Divide the dataset into training and testing data.
2. Initialize all weights.
3: While stopping condition is false, do step 4
4: For each training input:
Feed forward phase
i. Propagates inputs to nodes in the hidden layer.
ii. Each hidden unit sums its weighted inputs.
iii. Each output unit sums its weighted inputs.
iv. Applied its activation function to compute its output.
Backpropagation phase
v. Each output unit receives a target corresponding to the input training pattern, calculate error between actual and target output.
vi. Propagate error backward to adjust weights.
vii. Each output unit updates its bias and weights.
viii. Test stopping condition.
5: Apply this prepared model on testing data.

The following diagram shows flow of back propagation algorithm.

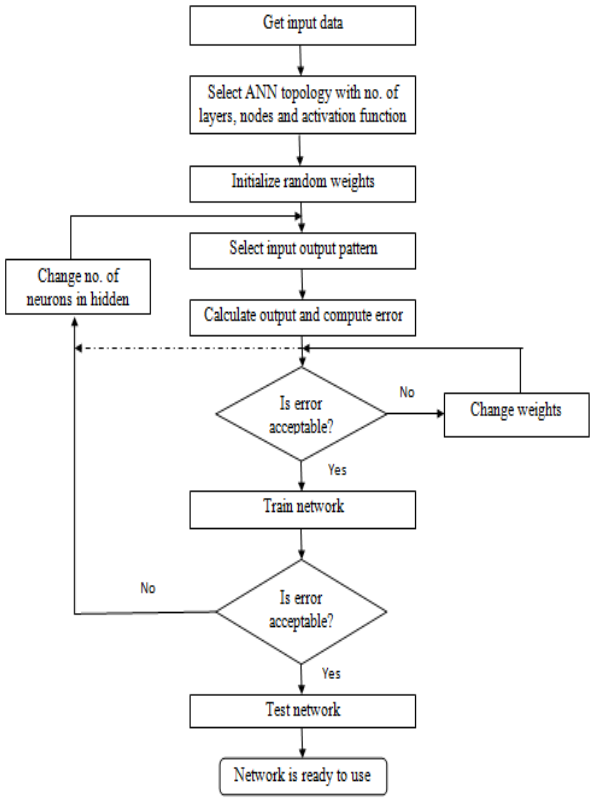

Figure 3: The Back Propagation (BP) Algorithm

IV. EXPERIMENTAL RESULTS

Dataset1: The germen credit dataset is used for fraud detection problem. This data is publicly available at UCI machine learning repository [4] [16]. It contains credit holders' information. This data has two types of credits like good and bad. There are total 24 inputs and one output.

Dataset2: This dataset is used for customer retention problem. This data is prepared under the guidance of bank. It contains bank customer's information such as customer id, age, gender, balance, income, credit card status, marital status, loan type, account type, number of transaction he makes, education and job of customer. It contains records of 1000 customers. It contains 12 inputs. This data has two types of customers like active and inactive.

Table 2: Dataset characteristics

Dataset	Input	Data set size	Training set size	Testing set size
Dataset1 (D1)	24	1000	700	300
Dataset2 (D2)	12	1000	700	300

This datasets are divided into training and testing set in in the ratio of 7:3, it means out of 1000 records 700 records are used for training and remaining 300 records for testing purpose and result is verified. On training data back propagation artificial neural network is applied and prepared model. Here logistic sigmoid activation function is used at hidden layer. After completion of training this model is used for testing data.

A simple thresholding method is used to divide data into classes. Here 0.5 threshold value is used to differentiate customer and credit group for two dataset. If output is greater than 0.5 then input belongs to one class (active for D2 and good for D2) otherwise it belongs to another class (inactive for D2 and bad for D2). This method works efficiently.

Figure 4 and 5 shows actual class and predicted class for some samples of testing data of dataset1 and dataset2 respectively.

```
> testoutputvspred = cbind(credittest$X1.7, testoutput)
> testoutputvspred

     Actual  Predicted
897       1          1
899       0          1
903       1          1
904       1          1
910       1          0
914       0          0
921       1          1
922       0          1
940       1          1
949       0          1
954       1          0
957       1          1
966       0          1
976       1          1
979       0          0
980       0          1
988       1          0
992       1          1
993       1          0
999       1          1
```

Figure 4: Testing samples for dataset1 with actual and predicted class

```
> testoutputvspred = cbind(test1$class, testout)
> testoutputvspred

    Actual  Predicted
106      1          1
110      1          1
116      0          0
119      1          1
126      1          0
127      0          0
173      1          1
573      1          1
576      1          1
582      0          0
583      0          1
584      1          1
587      0          0
588      0          0
848      1          1
995      1          1
```

Figure 5: Testing samples for dataset2 with actual and predicted class

The results are mentioned in table 3. It describes Root Mean Square Error (RMSE) and accuracy on training and testing datasets. From the results, it is clear that error for training stage is very less for both datasets. There is only 0.28 and 0.014 misclassification error for dataset 1 and dataset 2 respectively. The 72% and 98% accuracy is obtained for dataset1 and dataset2 respectively. It means out of 300 records it predicts correct class for about 295 records of dataset2. Hence using this model we can identify customers or credits status efficiently.

Table 3: Results on training and testing data

Dataset	Hidden nodes	RMSE		Accuracy %	
		Train	Test	Train	Test
D1	10	0.000003	0.0444	99	72
D2	7	0.000188	0.1093	99	98

V. CONCLUSION

In banking field massive volume of data is continuously generating. This data can be used to extract meaningful information from it. In this work we have used two datasets, bank customer's data and germen credit data to make classification. For classification purpose supervised artificial neural network is used. This algorithm gives 72% and 98% accuracy for dataset1 and dataset2 respectively. From results it shows that developed model works efficiently for two datasets.

REFERENCES

[1] Utkarsh Srivastava, Santosh Gopalkrishnan, "Impact of Big Data Analytics on Banking Sector: Learning for Indian Banks", ELSEVIER 2015.

[2] Amir E. Khandani, Adlar J. Kim, Andrew W. Lo, Consumer credit-risk models via machine-learning algorithms, ELSEVIER 2010.

[3] Francisca Nonyelum Ogwueleka, Department of Computer Science, Federal University of Technology, Minna "Neural Network and Classification Approach in Identifying Customer Behaviour in the Banking Sector: A Case Study of an International Bank", 2011.

[4] Iain Brown, Christophe Mues," An experimental comparison of classification algorithms for imbalanced credit scoring data sets", Expert Systems with Applications 39 (2012) 3446–3453, ELSEVIER.

[5] BenlanHea, Yong Shic, Qian Wan, Xi Zhao, "Prediction of customer attrition of commercial banks based on SVM Model", ELSEVIER 2014.

[6] HosseinHakimpoor, Islamic Azad University, Birjand Branch, Iran "Artificial Neural Networks' Applications in Management", World Applied Sciences Journal 14 (7), 2011.

[7] XinhuiTian, Rui Han, Lei Wang, Gang Lu, Jianfeng Zhan, "Latency critical big data computing in finance", ScienceDirect 2015.

[8] K. Chitra, B.Subashini, "Customer Retention in Banking Sector using Predictive Data Mining Technique", ICIT 2011.

[9] O. Oyeniyi A.B. Adeyemo, "Customer Churn Analysis in Banking Sector Using Data Mining Techniques", IEEE 2015.

[10] Dr. K. Chitra, B. Subashini, "Data Mining Techniques and its Applications in Banking Sector", IJETAE 2013.

[11] Dr. U. Devi Prasad Associate Professor Hyderabad Business School, GITAM University, Hyderabad "Prediction of Churn Behavior Of Bank Customers Using Data Mining Tools" 2012

[12] Kuchipudi Sravanthi, Tatireddy Subba Reddy, "Applications of Big data in Various Fields", IJCSIT 2015.

[13] N. Sun; J. G. Morris, J. Xu, X. Zhu, "iCARE: A framework for big data-based bankingcustomer analytics", IEEE 2014.

[14] XindongWu, Vipin Kumar, J. Ross Quinlan, JoydeepGhosh, Qiang Yang, "Top 10algorithms in data mining", Springer 2008.

[15] Mohammad Ridwan Ismail, Besut, Terengganu, Malaysia,"A Multi-Layer Perceptron Approach for Customer Churn Prediction", International Journal of Multimedia and Ubiquitous Engineering Vol.10, No.7 (2015).

[16] Cheng-Lung Huang, Mu-Chen Chen, Chieh-Jen Wang, "Credit scoring with a data mining approach based on support vector machines", ELSEVIER Expert Systems with Applications 33 (2007) 847–856.