ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijareeie.com

Vol. 6, Issue 4, April 2017

Review on Bank Data Classification

Vikram Kumar

Department of Computer Science and Engineering, Galgotias University, Yamuna Expressway Greater Noida, Uttar Pradesh, India

Email Id: dr.vikram12@rediffmail.com

ABSTRACT: Data mining is the mechanism by which significant rules are extracted from large and complicated data. Data mining throughout every field at the moment is increasing in popularity. Information units are being developed, in general, for consumer retention and recruitment in user-oriented industries such as advertising, finance and telecommunications. Data mining approaches utilize identification techniques to forecast potential customers in the relevant field in studies published for customer retention. In this analysis, bank advertisement data from the UCI Natural language processing Data Set have been used in individual data processing programs by computational modeling with the same labeling algorithms. Criteria of precision, precision and measurement were used to test the identification models ' efficiency. The research and training sets of data are separated arbitrarily by the holdout process for assessing the data set output while constructing identification models. With the 60-40 per cent, 75-25% and 80-20% differentiation ranges the data set was split into instruction and test sets. R, Knime, RapidMiner and WEKA are the data mining applications used in these systems. The k-nearest neighbor (k-nn), Naive Bays and C4.5 decision tree are also the categorization neural network commonly employed on these operating systems.

KEYWORDS*:* Data mining, Banking, Customer Acquisition, Data mining programs, Algorithms, Models.

I.INTRODUCTION

Data mining is widely used in many sectors of safety, economics and schooling to solve problems. Data mining experiments in customer-based business sectors such as telecom, insurance & banking to work on customers' shipment or customer satisfaction in the profession of quality of life for the treatment of the infection are being undertaken [1]. A forecast study was conducted to see whether a bank's initiative resulted in a recent acquisition of customers. Another aim of this research paper is to see deals with different data mining schemes with the same categorization algorithms. In the results section of the journal, the findings were shown in columns [2]. There is a large number of continuously constructed identification algorithms for different applications in banking literature. Some researchers have established consumers who have reacted positively to promotions using various methods including artificial neural networks via consumer segmentation [3]. In order to discover complex relationships between data, this paper has been using machine learning strategies of financial institutions and other organizations. In order for Keramati*et al.* to forecast existing clients, who would favor profitable banks with the telecommunications company's data-based in Iran, decision-making trees, information processing networks, closest neighbors and supporting vector machines have been used. The best result was described by attributing the methodologies used in the study [4–6].

II.OVERVIEW

The data mining and classification algorithms and data extraction systems used during the entire study are described in table 1 and 2.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijareeie.com

Vol. 6, Issue 4, April 2017

TABLE 1: classification algorithms

A. DATA MINING	Data mining is the process of extracting meaningful and structures information in the complex data sets. During this procedure, data mining methods such as classification, clustering and association rules are used.
B. CLASSIFICATION ALGORITHMS USED IN THE STUDY	In this study, bank marketing data set in UCI Machine Learning Data Set was used. Models were created using classification algorithms on this data set. Classification algorithms used in the study are the k-nearest neighbor (k-nn), Naive Bayes (NB), and C4.5 decision tree.
C. K-NEAREST NEIGHBOR ALGORTIHM (K-NN)	It is one of the most basic algorithms of sample-based learning algorithms. In this algorithm learning process is performed with the data in training set.
D. NAIVE BAYES ALGORITHM	The Naive Bayes algorithm is named after the English mathematician Thomas Bayes. Bayesian algorithms are among the statistical classification techniques and are based on the statistical Bayesian theorem. Bayes classifier is a predictive model, easier to apply.
E. C4.5 DECISION TREE ALGORTIHM	C4.5 algorithm has been developed by Ross Quinlan. The Gain Ratio is used in the C4.5 decision tree. C4.5 algorithm can work with either categorical or numerical attributes. Decision Trees generated by C4.5 can be used for classification.

TABLE 2: Data mining algorithms

A. DATA MINING PROGRAMS	**Numerous programs have been developed to implement data mining applications. Commercial programs such as SAS and open source programs such as RapidMiner (YALE), Waikato Environment for Knowledge Analysis (WEKA), R, Konstanz Information Miner (KNIME) can be given as examples of data mining programs developed**
B. KNIME	Konstanz Information Miner (Knime) is a data mining program developed by the Konstanz University data science team. Knime can import data of various file extensions (such as .txt, .arff, .csv)
C. RAPIDMINER (YALE)	It is a program developed by Ralf Klinkenberg, Ingo Mierswa and Simon Fischer in Artificial Intelligence Unit of Dortmund University of Technology. The Yale program has been developed at Yale University. Yale has been reintroduced in 2007 with the RapidMiner name

III.APPLICATION

In this analysis, an application has been made for data mining classified algorithms to forecast consumer purchases using the UCI registry marketing data. Data Set Software has been used by the development of models of the same

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijareeie.com

Vol. 6, Issue 4, April 2017

classified strategies in separate data mining systems for banking social media data set in UCI Natural language processing Data Set. The registry for banking marketing includes 18 attributes and 56322 customer documents [7]. Table 3 displays the type of information and assign description.

TABLE 3 BANK DATA SET

No	Attributes	Explanation of Attributes	Data Type
1	age	Customer's age	Numeric
2	job	Business status of the customers	Nominal
3	marital	Customer's marital status	Nominal
4	education	Customer's educational status	Nominal
5	default	Credit debt situation?	Nominal
6	balance	Average annual balance	Numeric
7	housing	Real estate debt situation?	Nominal
8	loan	Personal debt situation?	Nominal
9	contact	Type of communication	Nominal
10	day	Last interview day	Numeric
11	month	Last interview month	Numeric
12	duration	Last call duration	Numeric
13	campaign	Number of customers searching for the campaign during the campaign	Numeric
14	pdays	The number of times the customer has been called since the previous campaign	Numeric
15	Previous	How many times the customer is called before the campaign	Numeric
16	Poutcome	The end of the previous marketing campaign	Nominal
17	Customer	Is the customer a bank customer?	Nominal

- Evaluation parameters for product results

Various methods are used to evaluate the model generated by classification computer programs. The uncertainty matrix is one of these approaches [8]. Table 4 shows the current meanings and expected values for the classification algorithm. Table 4, explains the success appraisal requirements for classified applications.

Table 4: Appraisal requirements for classified applications

		Prediction	
		True	***False***
Actual	***True***	TT	TF
	False	FT	FF

Models are generated to measure the data set using classification algorithms. The classification mechanisms are split into preparation and relevant data to see the efficiency of the detection models.For this splitting procedure different methods have been created [9]. The holdout protocol was used in this analysis among these approaches. The experiment and thetestingdataframe was segregated once with a different proportion in the hold-out segretion. The flow of such a process is shown in figure 1.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijareeie.com

Vol. 6, Issue 4, April 2017

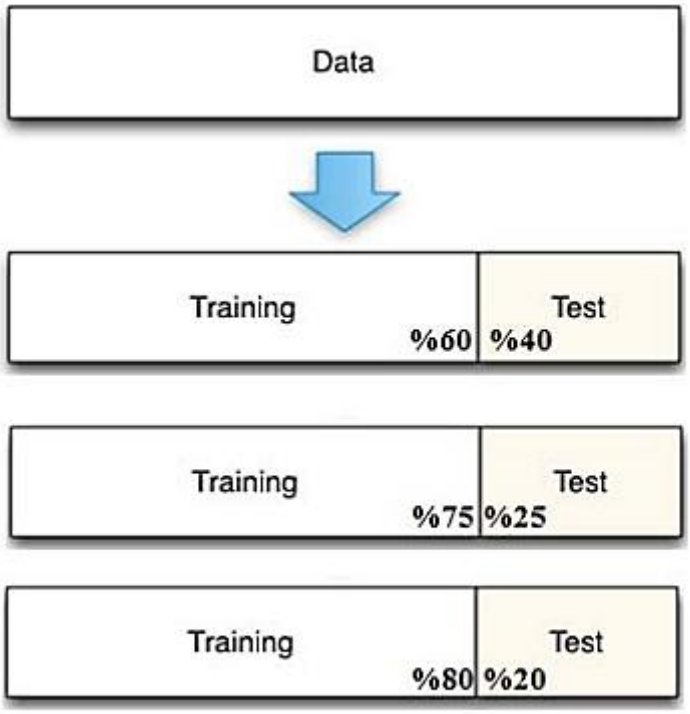

Fig. 1. Test and training set separation with the hold out method.

Formulations:
Table 4 consistency and error function is given by "Eq" (1) for the model created by the classification algorithms.

$$Accuracy = \frac{TT + \mathrm{FF}}{TT + TF + FT + FF}$$

$$Error = 1 - \mathrm{Accuracy} \qquad (1)$$

Pinpoint accuracy and responsiveness principles are offered by Eq (2) (3) for the model created by the Table 4 categorization machine learning.in each scenario.

$$Precision = \frac{DD}{DD + YD}$$

$$Sensitivity = \frac{TT}{TT + TF} \qquad (2)\ \&\ (3)$$

Equation shall specify the design consistency and F-measurement values of the identification formulas in Table 3. (4) and Eq, respectively. (5), in each case

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijareeie.com

Vol. 6, Issue 4, April 2017

$$Specificity = \frac{YY}{YY + YD}$$

$$F - measure = \frac{2 \times Sensitivity \times Precision}{Sensitivity + recision} \qquad (4)\ \&\ (5)$$

IV.RESULTS

The banking advertisement data set was used by "R, Knime, Weka, and RapidMiner data-gathering programs. C4.5 Tree of Decision and Bayes categorization neural networks" commonly reveal in those programs. Models with k-nearest neighbor were developed. The reliability, reliability and calculation parameters of these prototypes have been assessed. The training set and the assessments in each data gathering system were opposed to "60% to 40%, 75% to 25%, 80% to 20% and 90% to-10%" to check the efficiency of all the components. The analysis included a contrast between "60% and 40%". Such sections are listed including both in Table 5, Table 6 and Table 7.

Table 5: RESULTS OBTAINED IN R, KNIME, RAPIDMINER AND WEKA PROGRAMS WITH 60% HOLDOUT SEPARATION

Criteria	Accuracy			Precision			F-measure		
Algorithms	*NB*	*k-nn*	*C4.5*	*NB*	*k-nn*	*C4.5*	*NB*	*k-nn*	*C4.5*
R	0.872	0.871	**0.904**	0.930	0.909	**0.933**	0.927	0.928	**0.946**
Knime	0.866	0.860	**0.902**	0.921	0.867	**0.938**	0.825	0.900	**0,935**
RapidMiner	0.861	0.846	**0.885**	0.916	0.857	**0.932**	0.880	0.890	**0,918**
Weka	0.881	0.864	**0.900**	0.936	0.913	**0.940**	0.933	0.924	**0.944**

The same figures were reported for the success analysis in Table 5, Table 6, and Table 7. The "C4.5 decision tree was the best match in all aspects of the success metrics. In fact, the Weka system, whereas the R method produced better outcomes on the other three parameters, produced better performance".

Table 6 : RESULTS OBTAINED IN R, KNIME, RAPIDMINER AND WEKA PROGRAMS WITH 75% HOLDOUT SEPARATION

Criteria	Accuracy			Precision			F-measure		
Algorithms	*NB*	*k-nn*	*C4.5*	*NB*	*k-nn*	*C4.5*	*NB*	*k-nn*	*C4.5*
R	0.875	0.875	**0.906**	0.93	0.91	**0.935**	0.92	0.93	**0.948**
Knime	0.842	0.870	**0.905**	0.88	0.88	**0.933**	0.88	0.87	**0.90**
RapidMiner	0.869	0.842	**0.885**	0.88	0.88	**0.930**	0.86	0.86	**0.93**
Weka	0.882	0.865	**0.902**	0.93	0.91	**0.937**	0.93	0.92	**0.947**

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijareeie.com

Vol. 6, Issue 4, April 2017

Table 7: RESULTS OBTAINED IN R, KNIME, RAPIDMINER AND WEKA PROGRAMS WITH 80% HOLDOUT SEPARATION

Criteria	Accuracy			Precision			F-measure		
Algorithms	*NB*	*k-nn*	*C4.5*	*NB*	*k-nn*	*C4.5*	*NB*	*k-nn*	*C4.5*
R	0.875	0.87	0.906	0.934	0.911	**0.935**	0.928	0.93	0.947
Knime	0.842	0.87	**0.905**	0.882	0.880	**0.933**	0.883	0.87	**0.90**
RapidMiner	0.869	0.84	**0.88**	0.880	0.882	**0.930**	0.865	0.86	**0.93**
Weka	0.882	0.86	**0.902**	0.931	0.913	0.937	0.933	0.92	**0.946**

V.CONCLUSION

Data mining approaches utilize identification techniques to forecast potential customers in the relevant field in studies published for customer retention. In this analysis, bank advertisement data from the UCI Natural language processing Data Set have been used in individual data processing programs by computational modeling with the same labeling algorithms. Criteria of precision, precision and measurement were used to test the identification models ' efficiency. The research and training sets of data are separated arbitrarily by the holdout process for assessing the data set output while constructing identification models in this analysis, models of categorization machine learning were tested for the success of separate data-mining programs. In the four initiatives used, different findings were collected. Nevertheless, the decision tree optimization was the most efficient implementation in all programs. This result shows that the decision-tab approach provides better results independent of the system. Further studies are also needed to support this result by working with data other than the bank data set. This is a subject of another research to be further investigated in the future.

REFERENCES

[1]The World Bank, “Data: World Bank Country and Lending Groups,” World Bank Country and Lending Groups, 2018. .
[2]C. L. Huang, M. C. Chen, and C. J. Wang, “Credit scoring with a data mining approach based on support vector machines,” Expert Syst. Appl., 2007.
[3]E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, “The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature,” Decis. Support Syst., 2011.
[4]S. Agarwal, “Data mining: Data mining concepts and techniques,” in Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013, 2014.
[5]J. Tang, S. Alelyani, and H. Liu, “Feature selection for classification: A review,” in Data Classification: Algorithms and Applications, 2014.
[6]Y. Zheng, “Trajectory data mining: An overview,” ACM Transactions on Intelligent Systems and Technology. 2015.
[7]S. Wang and W. Shi, “Data mining and knowledge discovery,” in Springer Handbook of Geographic Information, 2012.
[8]G. Kesavaraj and S. Sukumaran, “A study on classification techniques in data mining,” in 2013 4th International Conference on Computing, Communications and Networking Technologies, ICCCNT 2013, 2013.
[9]C. Ratanamahatana and E. Keogh, “Everything you know about dynamic time warping is wrong,” Third Work. Min. Temporal Seq. Data, 2004.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijareeie.com

Vol. 6, Issue 4, April 2017

[10]P. Kaur, M. Singh, and G. S. Josan, "Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector," in Procedia Computer Science, 2015.
[11]W. Chen, G. Xiang, Y. Liu, and K. Wang, "Credit risk Evaluation by hybrid data mining technique," Syst. Eng. Procedia, 2012.
[12]C. Költringer and A. Dickinger, "Analyzing destination branding and image from online sources: A web content mining approach," J. Bus. Res., 2015.
[13]L. Dhanabal and S. P. Shantharajah, "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms," Int. J. Adv. Res. Comput. Commun.Eng., 2015.
•R Santhya, S Balamurugan, "A Survey on Privacy Preserving Data Publishing of Numerical Sensitive Data", International Journal of Innovative Research in Computer and Communication Engineering , Vol. 2, Issue 10, October 2014
•BalamuruganShanmugam, Dr.VisalakshiPalaniswami, Santhya. R, Venkatesh. R.S., "Strategies for Privacy Preserving Publishing of Functionally Dependent Sensitive Data: A State-of-the art Survey. Aust. J. Basic & Appl. Sci., 8(15): 353-365, 2014
•AnupamBaliyan, Vishal jain, Manish Kumar, Achin Jain and Uttam Singh, "Performance Analysis of Amdahl's and Gustafson's Law under multicore Processor Architecture", INDIACom-2018, 5th 2018 International Conference on "Computing for Sustainable Global Development", 14th – 16th March, 2018, held at BharatiVidyapeeth's Institute of Computer Applications and Management (BVICAM), New Delhi (INDIA).
•BishwajeetPandey, Vishal Jain, Rashmi Sharma, MragangYadav, D M Akbar Hussain, "Scaling of Supply Voltagein Design of Energy Saver FIR Filter on 28nm FPGA", International Journal of Control and Automation (IJCA), having ISSN No. 2005-4297, Vol. 10, No. 12, December, 2017, page no. 77 to 88.
•GauravVerma, Harsh Agarwal, Shreya Singh, ShaheemNighatKhinam, Prateek Kumar Gupta and Vishal Jain, " Design and Implementation of Router for NOC on FPGA", International Journal of Future Generation Communication and Networking (IJFGCN), Vol. 9, No. 12, December 2016 page no. 263 – 272 having ISSNo. 2233-7857 .