International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 9 • Issue 2 pp. 123-129 March 2018 - Sept 2018 www.csjournals.com

Selection of Classification Algorithm Based on K-Fold Cross Validation and Confusion Matrix(KFCM) For Direct Marketing

SACHIN JAIN[1*], DR. NARESH TRIVEDI[1,2]
[1]Department of Computer science and Engineering, Sunderdeep Engg College, Ghaziabad (U.P), India

Abstract: In identifying potential customers who would have requirement for a loan by using direct marketing, data mining techniques come to our rescue. In order to identify potential customers from a very large data we need an algorithm that optimizes two parameters, (i) high classification accuracy and (ii) minimum of error rates. In this paper we propose a Kernel-fold Based Confusion Matrix (KFCM) approach that when applied to existing Logistic regression, Random Forest, SVM, AdaBoost, Stochastic Gradient and Naive Bayes Data Mining Algorithms, narrows down the list of potential customers who may have requirements for loan. It has been observed that for Logistic Regression algorithm that there is significant improvement in classification accuracy. In this paper Data Set used is taken from UCI Machine Learning Repository.
Key words: Confusion Matrix, Kernel Logistic Regression, Random Forest, SVM, AdaBoost, Stochastic Gradient, Naive Bayes, Data Mining.

1. Introduction

Data Mining deals with access of information intelligently to support decision [1]. Data mining helps us in analyzing hidden patterns of data by utilizing associations, categorization, classification and clustering techniques. These hidden patterns facilitate business decision and help us to cut cost and increase profit [6]. Banking sector provides various services to customers regularly. One such service of banking that compiles and process information regarding potential loan customers. Bank may introduce its goods and services through advertising on TV, Radio, newspaper, internet, social media etc or by targeting potential customer directly through calls, mailers, bulk SMS etc [2]. Problem with direct marketing is that customer can at times feels disturbed or get offended and it can harm or downgrade rating of banks. It is essential to determine potential customer list carefully [3]-[5]. Over the time data to be processed by bank will grow, here role of data mining algorithm play a pivotal role in preparing or classifying potential loan customer data that serves to assist in decision making.

In this paper we will compare existing classification algorithms and determine which algorithm with our proposed method gives consistent or improved result to identify potential loan customers correctly so that cost of targeting loan customer can be reduced and profit of the financial institution or bank can be improved.

The remaining of the paper is structured as follows: Next section discusses the related work in brief, section III describes dataset, section IV describes our approach, section V describes proposed algorithm, section VI has discussion of results and analysis of various classification algorithms and section VII concludes the work.

2. Related Work

Related research that discusses the forecast about the potential customers to borrow loan is as follows:

Moro (2014), et al. [8] suggested a data-driven approach to determine the success of bank telemarketing by comparing four methods of data mining classification algorithms, i.e. Neural Networks, SVM and Logistic Regression, Decision Tree. Their paper highlighted that neural network algorithm gives the better result in comparison to other classification methods.

Singoei, L., J. Wang. 2013, et al. [9] proposed Data mining structure for direct marketing: A case study of bank marketing by Decision tree algorithm was chosen for classification and prediction. The original dataset was

arbitrarily partitioned into disjoint subsets. Few subsets are used to form the training set while the leftover subsets are used for the testing set. Decision tree generate two classes, the positive class and negative class based on the responders and non-responders. Positive class set gives better results for target marketing.

Shamala Palaniappan, Aida Mustapha, Cik Feresa Mohd Fooz Rodziah Atan, et al. [10] proposed Customer Profiling using Classification approach for Bank Telemarketing by comparing classification algorithms Random Forest, Naïve Bayes and Decision Tree. This Paper determines accuracy, precision and recall rates and justifies that decision tree classification generates better accuracy while random forest classification generates better precision to determine potential customer profiles and improve telemarketing sales.

Swati Jadhav, Hongmei He, Karl Jenkins et al. [11] proposed an Academic review: applications of data mining techniques in finance industry and conclude that data mining techniques like Decision trees and Neural Networks produces better results than other methods such as SVM, Regression and Hybrid models, Markov model, Fuzzy set theory, KNN and Association Rule Mining.

Various classification algorithms classify data in different ways. Multiple classification algorithms exist in Data Mining but which algorithm will generate consistent result is still a problem. Our proposed method will help financial Institutions and Banks to determine a classification algorithm that produces better results in terms of classification accuracy. Proposed method is based on K fold value (K may have value varies from 5 to 20). Classification Algorithm that generates consistent or improve classification accuracy result on all K folds value is considered as the best algorithm to classify potential loan customer data.

3. Data Set Used

The dataset used in this paper is taken from UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/bank+marketing#)[7]. Data set have 41188 attributes. All Attributes have no missing Values. Here 66% data is used as Training Data, and remaining 34% data as Test data, to determine the accuracy of classification.

4. Our Approach

Here our approach is to calculate Confusion matrix of different classification algorithms based on different K folds Cross Validations and then arrive at a better algorithm as per needs of bank based on target attribute Y namely the Classification accuracy. Firstly we apply it on Logistic Regression classification techniques than we will compare it with other algorithms like Random Forest, SVM, AdaBoost, Stochastic Gradient, Naive Bayes.

5. Proposed Algorithm(KFCM)

Step1: For (k= Min to Max) // Min value for k=5 and Max Value for k=20 // Where K is different fold Cross Validations

To classify data, we will use ordered k-fold cross-validation, in which the folds are chosen and each fold has nearly the equal number of class labels.

In k-fold cross-validation, the original data is arbitrarily divided into k equal size subsamples. Any one subsample is chosen as the validation data for testing, out of the available subsample, and the remaining subsamples can be called as training data. The cross-validation process is then looped k times, with each of the k subsamples processed once only.

Step2: To Calculate Confusion Matrix of different Algorithm and find classification accuracy based on given formula.

Performance of a classification algorithm can be determined by a confusion matrix table on a set of test data of known true values.

Performance of a classification algorithm can be determine with help of a confusion matrix, which is a table on set a set of test data defined as

	Class 1	**Class 2**
Class 1 Actual	True Positive	False Negative
Class 2 Actual	False Positive	True Negative

Class 1: Positive Class 2: Negative

Where:
P:-Result is positive
N:-Result is not positive
True Positive (TP) :- A true positive test outcome detects the condition when the condition is present.
False Negative (FN) : A false negative test outcome does not detect the condition when the condition is present.
True Negative (TN) : A true negative test outcome does not detect the condition when the condition is absent
False Positive (FP) : A false positive test outcome detects the condition when the condition is absent.

Classification Accuracy:
It is calculated as :

$$\textbf{Classification Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}}$$

Step3: To Calculate Classification Accuracy, F1, Precision, Recall Value on the basis of Confusion Matrix. Recall can be defined as:

Recall:
High Recall determines that the class is accurately determined (less number of False Negative).
Recall can be determined by the following expression:

$$\textbf{Recall} = \frac{TP}{TP + FN}$$

Precision:
Value of precision can be determined by the division of accurately classified true positive example by the all the positive example.

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

High recall, low precision: This indicates that most of the positive examples are correctly identified but it have lot of false positives.
Low recall, high precision: This indicates that we miss lots of positive examples (high False Negative) but those we identified as positive are actually positive (low False Positive)

F-measure can be defined as:
F-measure:
In calculation of F-measure it uses Harmonic Mean inplace of Arithmetic Mean as it discards extreme values.
The F-Measure used to be closer to the lesser value of Precision or Recall.

$$\textbf{F- Measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

Step4: Always select the Classification algorithm based on consistent and high classification accuracy with varying K-Fold Value.

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 9 • Issue 2 pp. 123-129 March 2018 - Sept 2018 www.csjournals.com

Result and Analysis (k Fold=5)

	0	**1**	**Σ**
0	35615	933	36548
1	2761	1879	4640
Σ	38376	2812	41188

Table 1. Confusion Matrix Using Logistic Regression

	0	**1**	**Σ**
0	30816	5732	36548
1	1614	3026	4640
Σ	32430	8758	41188

Table 2 .Confusion Matrix Using Naive Bayes

	0	1	Σ
0	35307	1241	36548
1	2515	2125	4640
Σ	37822	3366	41188

Table 3. Confussion Matrix Using Random Forest

	0	1	Σ
0	35091	1457	36548
1	2612	2028	4640
Σ	37703	3485	41188

Table 4. Confussion Matrix Using Stochastic Gradient

	0	1	Σ
0	32449	4099	36548
1	3670	970	4640
Σ	36119	5069	41188

Table 5. Confussion Matrix Using SVM

	0	1	Σ
0	34234	2314	36548
1	2278	2362	4640
Σ	36512	4676	41188

Table 6. Confussion Matrix Using ADA Boost

Method Name	**Classification Accuracy**	**F1**	**Precision**	**Recall**
Logistic Regression	0.910	0.900	0.899	0.910
Naive Bayes	0.822	0.844	0.882	0.822
Random Forest	0.909	0.902	0.899	0.909
Stochastic Gradient	0.901	0.895	0.891	0.901
SVM	0.811	0.815	0.819	0.811
AdaBoost	0.889	0.889	0.889	0.889

Table 7.

All the Analysis is done on Number of Folds =5, Repeat Train/ test=10, Training Set Size =66%.

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 9 • Issue 2 pp. 123-129 March 2018 - Sept 2018 www.csjournals.com

Result and Analysis (k Fold=10)

	0	**1**	**Σ**
0	35559	949	36548
1	2765	1875	4640
Σ	38364	2824	41188

Table 8. Confusion Matrix Using Logistic Regression

	0	**1**	**Σ**
0	30842	5706	36548
1	1614	3026	4640
Σ	32456	8732	41188

Table 9 .Confusion Matrix Using Naive Bayes

	0	1	Σ
0	35300	1248	36548
1	2493	2147	4640
Σ	37793	3395	41188

Table 10. Confussion Matrix Using Random Forest

	0	1	Σ
0	35294	1254	36548
1	2863	1777	4640
Σ	38157	3031	41188

Table 11. Confussion Matrix Using Stochastic Gradient

	0	1	Σ
0	32489	4059	36548
1	3694	946	4640
Σ	36183	5005	41188

Table 12. Confussion Matrix Using SVM

	0	1	Σ
0	34155	2393	36548
1	2195	2445	4640
Σ	36350	4838	41188

Table 13. Confussion Matrix Using ADA Boost

Method Name	**Classification Accuracy**	**F1**	**Precision**	**Recall**
Logistic Regression	0.910	0.900	0.898	0.910
Naive Bayes	0.822	0.844	0.882	0.822
Random Forest	0.909	0.901	0.899	0.909
Stochastic Gradient	0.900	0.891	0.887	0.900
SVM	0.812	0.815	0.818	0.812
AdaBoost	0.889	0.890	0.891	0.889

Table 14.

All the Analysis is done on Number of Folds =10, Repeat Train/ test=10, Training Set Size =66%.

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 9 • Issue 2 pp. 123-129 March 2018 - Sept 2018 www.csjournals.com

Result and Analysis (k Fold=20):

	0	**1**	**Σ**
0	35610	938	36548
1	2769	1871	4640
Σ	38379	2809	41188

Table 15. Confussion Matrix Using Logistic Regression

	0	**1**	**Σ**
0	30839	5709	36548
1	1616	3024	4640
Σ	32455	8733	41188

Table 16. Confussion Matrix Using Naive Bayes

	0	1	**Σ**
0	35349	1199	36548
1	2518	2122	4640
Σ	37867	3321	41188

Table 17. Confussion Matrix Using Random Forest

	0	1	**Σ**
0	35340	1208	36548
1	2890	1750	4640
Σ	38230	2958	41188

Table 18. Confussion Matrix Using Stochastic Gradient

	0	1	**Σ**
0	33406	3142	36548
1	3805	835	4640
Σ	37211	3977	41188

Table 19. Confussion Matrix Using SVM

	0	1	**Σ**
0	34195	2353	36548
1	2235	2405	4640
Σ	36430	4758	41188

Table 20. Confussion Matrix Using ADA Boost

Method Name	**Classification Accuracy**	**F1**	**Precision**	**Recall**
Logistic Regression	0.910	0.900	0.898	0.910
Naive Bayes	0.822	0.844	0.882	0.822
Random Forest	0.910	0.903	0.900	0.910
Stochastic Gradient	0.901	0.891	0.887	0.901
SVM	0.831	0.826	0.820	0.831
AdaBoost	0.889	0.889	0.890	0.889

Table 21.

All the Analysis is done on Number of Folds =20, Repeat Train/ test=10, Training Set Size =66%.

6. Conclusion

We applied different techniques of data mining on banking data with and without KFCM based approach. We observe from sets given in Table 7, Table 14 and Table 21 that KFCM based logistic regression gives consistent Classification Accuracy as 91.0 %. However, for data mining algorithms Logistic Regression, Stochastic Gradient classification accuracy does not improve with F = 20 but post lower value as compared to Logistic Regression algorithm. For logistic Regression method with K-Fold 20, we can observe that out of 41188 cases only 4640 customers are eligible for loan. Finally we conclude that Logistic Regression with KFCM based analysis shows better performance in terms of better and consistent classification accuracy amongst all these algorithms.

References

[1] S. Abbas, "Deposit subscribe Prediction using Data Mining Techniques based Real Marketing Dataset," Int. J. Comput. Appl., vol. 110, no. 3, pp. 975-887, 2015.

[2] S. Moro and R. M. S. Laureano, "Using Data Mining for Bank Direct Marketing: An application of the CRISP-DM methodology," Eur. Simul. Model. Conf., no. Tableure 1, pp. 117–121, 2011.

[3] C. Vajiramedhin and A. Suebsing, "Feature Selection with Data Balancing for Prediction of Bank Telemarketing," Appl. Math. Sci., vol. 8, no. 114, pp. 5667–5672, 2014.

[4] H. A. Elsalamony, "Bank Direct Marketing Analysis of Data Mining Techniques," Int. J. Comput. Appl., vol. 85, no. 7, pp. 12–22, 2014.

[5] H. Elsalamony and A. Elsayad, "Bank Direct Marketing Based on Neural Network," Int. J. Eng. Adv. Technol., vol. 2, no. 6, pp. 392–400, 2013.

[6] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, 2006.

[7] https://archive.ics.uci.edu/ml/datasets/bank+marketing#

[8] Moro, S., Laureano, R., and Cortez, P. 2011, "Using data mining for bank direct marketing: An application of the crisp-dm methodology" In Proceedings of European Simulation and Modelling Conference-ESM'2011 (pp. 117–121), Eurosis.

[9] Singoei, L., and J. Wang. 2013. Data mining framework for direct marketing: A case study of bank marketing. International Journal of Computer Science Issues (IJCSI) 10(2): 198–203.

[10] Shamala Palaniappan , Aida Mustapha , Cik Feresa Mohd Foozy, Rodziah Atan, "Customer Profiling using Classification Approach for Bank Telemarketing" International Journal On Informatics Visualization Vol 1 (2017) NO 4 – 2 e-ISSN : 2549-9904 ISSN : 2549-9610

[11] Swati Jadhav, Hongmei He, Karl Jenkins, "An Academic Review: Applications Of Data Mining Techniques In Finance Industry, International Journal Of Soft Computing And Artificial Intelligence", ISSN: 2321-404X, Volume-4, Issue-1, May-2016