International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in

Vol.3., Issue.2, 2015

RESEARCH ARTICLE

ISSN: 2321-7758

IMPLEMENTATION OF CLASSIFIERS AND THEIR PERFORMANCE EVALUATION

ANURAG SHRIVASTAVA[1], BHAVANA KUMARI[2]

[1]M.Tech Scholar, ACE&IT, Jaipur, Affiliated to RTU Kota

[2]M.Tech Scholar ACE&IT, Jaipur, Affiliated to RTU Kota

Article Received: 07/03/2015 | Article Revised on; 15 /03/2015 | Article Accepted on:18/03/2015

ANURAG SHRIVASTAVA

BHAVANA KUMARI

ABSTRACT

Data mining technique are used in various fields such as medical, business strategy, online marketing, bioinformatics, and whether forecasting. Classification technique is used to make classification models for the data mining techniques. This technique generally used for prediction, the matter is that how much accurate result they will predict. In proposed work evaluates the performance of various classifiers with and without using feature selection method, using various measures such as sensitivity, specificity and accuracy. For evaluating the performance of various classifiers used bank's direct marketing dataset and various classification methods are used. Finally it is concluded that Logistic Regression classifier is better than all Classifier with accuracy 90.03%.

Key words: - Data mining Technique, Classifiers, Nearest Neighbors, Naïve Bayes, Logistic regression, Feature Selection Method

INTRODUCTION

1.1 DATA MINING

Data mining is a technique by which fetch anonymous information from huge dataset, Data mining plays important role in knowledge discovery in dataset, knowledge discovery is a process by which converting useless data in to useful information, Data mining is new emerging evolutionary approach in field of technology It can be called as "knowledge mining from data" [1].

1.1.1 KD Process

Knowledge discovery is a process which shown in figure1.1 with various steps such as data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, and Knowledge presentation.

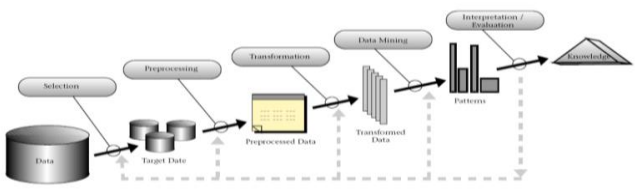

Fig:-1.1 Knowledge discovery as a process [1]

In given diagram there are various steps of knowledge discovery process in which dataset pass through these steps which are very needful for searching useful information from huge dataset, the first four steps are used before mining of data, the first step shows that a large dataset has various

types of useless contents such as noisy data, data redundancy, inconsistency and many more that has to be removed before mining. In second step various data sources are combined in single source. in third step there is selection of relevant data for mining ,in forth step data transformed in various forms for performing data mining operations ,after these steps data become ready for mining process.

1.2 CLASSIFICATION

This is most important technique of datamining to make various classification models for a given dataset [2]. The classification techniques are used to make models which are useful for prediction of future test dataset [3, 4]. The designing of Classification model defines a set of pre-determined classes.

Now Days, data mining techniques are being used by many industries including banking and finance. The bank's marketing department use data mining technique to analyze given customer datasets and prepare statistical profile of individual customer preference to product and little extra service.

1.3 FEATURE SELECTION ALGORITHM WITH CLASSIFICATION METHOD

A dataset consists of many features, some of them are useful and some are useless. The useless features take unnecessary extra time for their execution, and don't put any positive effect on prediction result, so it becomes necessary to select only those features which are useful for prediction to give more accurate result. Now a days the feature selection methods plays an evolutionary approach in the field of bio-informatics, whether forecasting, business strategies, online marketing campaign and banking field too, because this method provides reliable, more accurate results. In proposed work feature selection method with classifiers is used over the bank direct marketing dataset [5].

1.4.1 VALIDATIONS USED

1.4.1 Hold-out Validation

This is one of the easiest methods of validation method, in which a given dataset is divided in two parts, one is known as training dataset and other is testing dataset. It uses a function approximator to set the function values for training dataset and then it predicts output values for testing dataset.

In Proposed work hold out method is used which is easy to implement and better than cross validation because it does not produce complexity like cross validation. The cross validation calculate average result value of all experiments so it makes the model difficult, but the hold out method calculate result using training dataset over the test dataset.

1.5 DATA SET

Proposed work extracted the datasets of bank direct marketing from UCI repository. It has dimensions of 16 attribute and 45,211 instances. For proposes of training and testing, only 60% of the overall data uses for training and the remaining 40% dataset uses for testing the accuracy of the selected classification algorithms.

In proposed work the used dataset taken from bank direct marketing campaigns of Portuguese banking institute. The phone calls was the basic attributes for the bank marketing campaigns, it was compulsory for all clients to had more than one contact and asked to subscribe to the product(bank term deposit) or not [6]. The classifier model will predict that how many clients has subscribed a term deposit or not using the variable y. The bank direct marketing has 45211 observations based on 16 attributes/features.

1.6 PROPOSED WORK OBJECTIVES

The objective of the proposed work is to evaluate performance of different classifiers with and without using of feature selection method. Proposed work shows comparison of predicted values with actual data and show results in the form of confusion matrix. With the help of confusion matrix calculate various measures such as Accuracy, Sensitivity and Specificity, and perform comparison on the basis of these measures. For the classification following techniques are mainly used with and without feature selection method.

- Logistic Regression
- Nearest Neighbors
- Naive-Bayes
- Logistic Regression with Feature Selection
- Nearest Neighbors with Feature Selection
- Naive-Bayes with Feature Selection

In First scheme, Before classification, data preprocessing techniques such as data cleaning, is applied and In Second Scheme , data preprocessing techniques as well as feature selection technique is applied these techniques usually increase the efficiency of the algorithm for classifying the data correctly and find out which classifier predict result

with maximum accuracy. Analyze the performance of different classification techniques to select the one of the most accurate results for classification of bank's direct marketing dataset. [7, 8]

2. METHODOLOGY

In proposed work three classifiers have been implemented first with and then without using feature selection method and their performance analysis is done with comparison with each other, and finally it is concluded that which one shows best results. Three classifiers Logistic Regression Naïve Bayes and K-Nearest Neighbor with and without using feature selection method in matlab. For evaluation of performance of classifier, the dataset is splited in two parts one is training dataset and another is testing dataset with the hold out validation method.

Classifiers learn from training dataset and perform prediction on test dataset in the form of confusion matrix, which is a source to calculate three performance measures which are accuracy, sensitivity and specificity. In proposed work learning process divided in two schemes.

(i) First scheme is combination of:-

- Data preprocessing
- Learning algorithm

(ii) Second scheme is combination of:-

- Data preprocessing
- Feature Selection
- Learning algorithm

2.1 OVERVIEW OF THE LEARNING SCHEME USED IN PROPOSED WORK

2.1.1 First Learning Scheme

In Fig2.1 shown first learning scheme bank direct marketing dataset is divided in two parts first is training data set while other is testing dataset now both datasets passes through data preprocessing ,now different classifiers will trained from trainee dataset and prediction for the test dataset .

Classifier predicts the result in the form of confusion matrix; now performance will be evaluated using various measures such as Accuracy, Sensitivity, and Specificity.

2.1.2 Second Learning Scheme

In Second Learning Scheme the classifier is used with feature selection method. Feature selection method is used after data preprocessing step in which redundant, irrelevant, erroneous data and missing data are removed. Fig. 2.2 contains the details. After passing through feature selection data set goes as input to classifiers for learning.

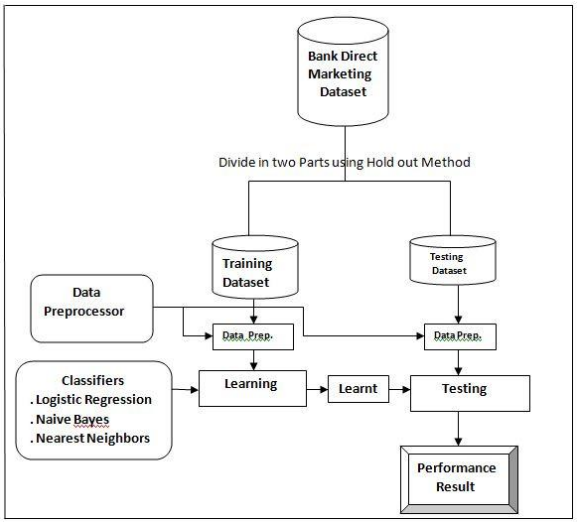

Fig 2.1:- First Learning Scheme

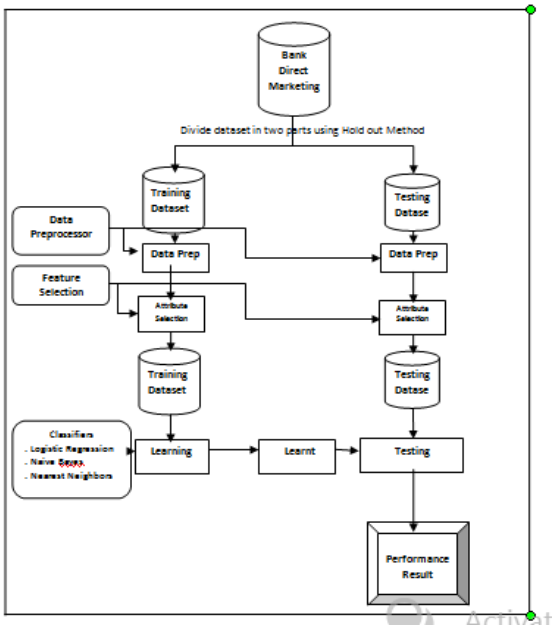

Fig 2.2:-Second Learning Scheme

The main obstacle was that how to divide bank direct marketing dataset into training dataset and test dataset, as above given scheme learning process does not depend on the test dataset. It is pre requisite condition for evaluation the performance of classifier for test dataset. For this Hold out method is used, it estimate that with how much accuracy a classifier will predict. It involves partitioned of dataset in to 60:40 ratio, refer 60% dataset for training and 40% dataset to training dataset.

2.2 TRAINING OF CLASSIFIER

After the division of trainee and test dataset in the ratio of 60 and 40, learning scheme with trainee dataset are used to construct to learner. A Second learning scheme consists of a data preprocessing method, a featuring selection method, and a learning algorithm. In first learning scheme which did not consider featuring selection method the detailed learner construction procedure is as follows:

i. **Data preprocessing-** Data preprocessing is an important part of construction of a learner. In this step, the training dataset are passed through various tiny processes, such as discrediting or transforming numeric attributes, and removing outliers, handling missing values.

ii. **Feature selection** Feature selection is a method which is also known as variable selection, attribute selection or variable subset selection; it is the process of selecting a subset of features which are useful in model construction.

iii. **Learner construction-** After the completion of attribute selection process, the training dataset contains best attributes which produce positive effect for learner construction then after filtered training dataset which have dimensionally reduced and classification method are used for construct to learner, after learning process the predicted result over the test dataset, that will be compared with actual values for evaluation their performance.

2.3 PREDICTION

The classifier, which is trained from trainee dataset, is then used to make a prediction on the test dataset. Predicted values will be compared with actual values and show the result in the form of confusion matrix. Confusion matrix is one of the function which is used for analyze the performance of a machine learning techniques.

Proposed work the performance of different classification techniques analyzed to select the one with the most accurate results for classification of bank direct marketing dataset. Three very commonly used techniques from different classification techniques are chosen from machine learning. In machine learning use three classifiers one is nearest neighbors, naive bayes and logistic regression.

2.3.1 Naïve Bayes classifier

Naïve Bayes classifier can learn and fetch useful information from hidden pattern in trainee dataset very efficiently with supervised learning setting .The main advantage of naïve bayes classifier is needed less training data to estimation of parameters necessary for classification; it is one of the most popular classifier in the datamining industry.

2.3.2 Logistic Regression

Logistic is one of the statically method for analyze to a dataset which contains more one independent variable which are used for determine to outcome. The outcome variable is measured by dichotomous variable; these are those variables which contain two possible outcomes.

2.3.3 K- Nearest Neighbors

K-Nearest neighbor is the instance based learning method it is designed by Mitchell in 1997.in instance based learning a distance function is used for analyze that which instance of training set is near by a new unknown instance. It means that this classifier take less time for learning but it takes more time for solving to a query.

2.3.4 Data set description

Proposed work extracted the datasets of bank direct marketing from UCI repository. It has dimensions of 16 attribute and 45,211 instances. For proposes of training and testing, only 60% of the overall data uses for training and the remaining 40% dataset uses for testing the accuracy of the selected classification algorithms.

2.3.5 Confusion Matrix

Confusion matrix has the information about actual and predicted values during classification process done by classifiers. It shows the result in the form of matrix, contains four elements which are True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). The difference between the predicted values and actual values shows in percentage of correctness/ incorrectness of classification. True Positive (TP) is the number of true prediction which is done classifier same as actual value for true instance, True Negative (TN) is number of true prediction for an instance which is false, now False Positive and False Negative occurs when false prediction done by classifier, False Positive (FP) is number of false prediction for the

positive instance and is not same as actual value, False Negative (FN) is number of false prediction for an instance which is false. Table 4.1 shows the confusion matrix for a two-class classifier.[9]

TABLE 2.1 CONFUSION MATRIXES

		Predicted Class	
		Cm1	Cm2
Actual Class	Cm1	True positives(TP)	False negatives(FN)
	Cm2	False positives(FP)	True negatives(TN)

2.4 Three Measures for performance

Accuracy- Accuracy of Classification is defined as the ratio of the number of correctly classified cases and is equal to the sum of TP and TN divided by the total number of cases N.[9]

$$\text{Acurracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

Sensitivity- Sensitivity measures the correctness of the predicted model. It is defined as the percentage of classes correctly predicted to be fault prone.[9]

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Specificity- Specificity also measures the correctness of the predicted model. It is defined as the percentage of classes predicted that will not be faulted prone.[9]

$$\text{Specificity} = \frac{TP}{TP + FP}$$

3. THE EXPERIMENTAL RESULTS ANALYSIS

After implementing and executing the proposed methodology the result obtained are as follows:

As previously stated, the first 16 attributes are defined as input attributes and the output attribute (y) is defined as a target.

Now perform experiment, the input for classifier is 16 attributes of dataset and the output attribute is y in which classifier has to predict that how many people has subscribed fixed deposit (yes) or (no), which has to be predict to classifier. In given dataset the actual values for y mean number of no is 39922 and number of y is 5289.which shown in following table 3.1.

TABLE 3.1DATASET VALUES FOR ATTRIBUTE Y

Value	Count	Percent
no	39922	88.30%
yes	5289	11.70%

First step is data preprocessing in which data is divided in two parts one part is trainee data set and other part is test data set, it will learn from trainee dataset that what is attributes values for who has subscribed the term deposit (yes) or not (no).In trainee dataset number of yes is 3179 and number of no is 23948, shown in following table 3.2.

TABLE 3.2DIVIDED TRAINEE DATASET VALUE OF Y

No	23948	88.28%
yes	3179	11.72%

Now classifier whatever learn, perform prediction for the attribute y that who has subscribed the term deposit (yes) or no on the test Dataset in which value of y attributes is 15947 for no and 2110 for yes, shown in table 3.3.

TABLE 3.3ACTUAL TEST DATASET VALUES FOR Y

Value	Count	Percent
no	15974	88.33%
yes	2110	11.67%

Now we compare prediction output which is predicted by different classifier with actual output shown in table 3.4, and evaluate accuracy, sensitivity and specificity of classifier using confusion matrix.

3.1 CLASSIFIER'S PREDICTION

3.1.1 Logistic Regression

Table 3.4 shows confusion matrix generated by

	C1	C2
C1	15579	395
C2	1368	742

Logistic Regression.

TABLE 3.4 THE CONFUSION MATRIX FOR LOGISTIC REGRESSION

Accuracy of Logistic Regression Classifier = 90.25
Sensitivity of Logistic Regression Classifier = 97.52
Specificity of Logistic Regression Classifier = 91.92

3.1.2 Logistic Regression after Feature Selection

TABLE 3.5CONFUSION MATRIX FOR LOGISTIC REGRESSION AFTER FEATURE SELECTION

Predicted class

Actual class	C1	C2
C1	15648	326
C2	1662	448

Accuracy of Logistic Regression Classifier after Feature Selection = 89.00
Sensitivity of Logistic Regression Classifier after Feature Selection = 97.95
Specificity of Logistic Regression Classifier after Feature Selection = 90.39

3.1.3 Nearest Neighbors:

TABLE 3.6 CONFUSION MATRIX FOR NEAREST NEIGHBORS

Predicted class

Actual class	C1	C2
C1	14994	980
C2	1311	799

Accuracy of Nearest Neighbors Classifier = 87.33
Sensitivity of Nearest Neighbors Classifier = 93.86
Specificity of Nearest Neighbors Classifier = 91.95

3.1.4 Nearest Neighbors after Feature Selection:

TABLE 3.7 CONFUSION MATRIX FOR NEAREST NEIGHBORS AFTER FEATURE SELECTION

Predicted class

Actual class	C1	C2
C1	14720	1254
C2	1424	686

Accuracy of Nearest Neighbors Classifier after Feature Selection = 85.19
Sensitivity of Nearest Neighbors Classifier after Feature Selection = 92.14
Specificity of Nearest Neighbors Classifier after Feature Selection = 91.17

3.1.5 Naive Bayes:

TABLE 3.8 CONFUSION MATRIX FOR NAIVE BAYES CLASSIFIER

Predicted class

Actual class	C1	C2
C1	14699	1275
C2	976	1134

Accuracy of Naive Bayes Classifier = 87.55
Sensitivity of Naive Bayes Classifier = 92.01
Specificity of Naive Bayes Classifier = 93.77

3.1.6 Naïve Bayes After feature selection

TABLE 3.9CONFUSION MATRIX FOR NAÏVE BAYES AFTER FEATURE SELECTION

Predicted class

Actual class	C1	C2
C1	14157	1817
C2	1141	969

Accuracy of Naive Bayes Classifier after Feature Selection = 83.64
Sensitivity of Naive Bayes Classifier after Feature Selection = 88.62
Specificity of Naive Bayes Classifier after Feature Selection = 92.54

3.2 COMPLETE RESULT

Table 3.10 shows comparison of classifiers performance in tabular form.

TABLE 3.10 COMPARISON OF CLASSIFIER PERFORMANCE

Classifiers	Performance Measures		
	Accuracy	Sensitivity	Specificity
Logistic Regression	90.25	97.52	91.92
Nearest Neighbors	87.33	98.60	91.95
Naïve Bayes	87.55	92.01	93.77

Table 3.11 shows comparison of classifier performance using feature selection method in tabular form.

TABLE 3.11: COMPARISON OF CLASSIFIER PERFORMANCE USING FEATURE SELECTION

Classifiers	Performance Measures		
	Accuracy	Sensitivity	Specificity
Logistic Regression with Feature Selection	89.00	97.95	90.39
Nearest Neighbors with Feature Selection	85.19	92.14	91.17
Naïve Bayes with Feature Selection	87.64	92.01	93.77

3.3 RESULT ANALYSIS

From table 3.10 and 3.11 it is clear that Logistic Regression shows highest accuracy 90.25%, Nearest Neighbor shows highest sensitivity 98.60% and Naïve Bayes shows highest specificity 93.77%.

4.5 COMPARATIVE ANALYSIS

When comparison of performance of classifiers in proposed work is done with previous work, it is found that classifiers used in proposed work gives maximum accuracy 90.25%, maximum sensitivity 98.60%,and maximum specificity 93.77% ,which shows an improvement of 3.3% in accuracy ,11.60% improvement in sensitivity and 7.07% improvement in specificity.

5. CONCLUSION

In Proposed work evaluation and comparison of the classification performance of three different data mining techniques Logistic Regression, Naïve Bayes and Nearest Neighbors with and without using feature selection on the bank direct marketing dataset to classify for bank term deposit subscription in binary form yes and no. For The evaluation of classification performances of the three techniques have been using three measures such as accuracy, sensitivity and specificity. This dataset is divided into parts one is training dataset and other is test dataset by the ratio 60% and 40%, respectively. Results of Experiments have shown the effectiveness of models. Logistic Regression has achieved 90.25% accuracy which is better than Naïve Bayes and Nearest Neighbors. And also as compared to previous work the proposed work gives an improvement of 3.3 % in accuracy, improvements of 11.60% in sensitivity and 7.07% in specificity.

REFERENCES

[1]. Eniafe Festus Ayetiran, "A Data Mining-Based Response Model for Target Selection in Direct Marketing", I.J. Information Technology and Computer Science, 2012, 1, 9-18.

[2]. SERHAT ÖZEKES and A.YILMAZ ÇAMURCU:" CLASSIFICATION AND PREDICTION IN A DATA MINING APPLICATION "Journal of Marmara for Pure and Applied Sciences, 18 (2002) 159-174 Marmara University, Printed in Turkey.

[3]. Kaushik H, Raviya Biren and Gajjar,"Performance Evaluation of Different Data Mining Classification Algorithm Using WEKA", Indian Journal of Research (PARIPEX) Volume: 2 | Issue: 1 | January 2013 ISSN - 2250-1991.

[4]. Wai Ho, Keith, C.C.Chan, Xin yao "A Novel Evolutionary Data Mining Algorithm with Applications to Churn Prediction", IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, Vol. 7, No. 6, Dec 2003, PP: 532- 545.

[5]. C. Akalya devi1, K. E. Kannammal2 and B. Surendiran3,M.E (CSE), Sri Shakhty Institute of Engineering and Technology, Coimbatore, India" HYBRID FEATURE SELECTION MODEL FOR SOFTWARE FAULT PREDICTION", International Journal on Computational Sciences & Applications (IJCSA) Vo2, No.2, April 2012.

[6]. [Moro et al., 2011] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation.

[7]. Reza Allahyari Soeini and Keyvan Vahidy Rodpysh: "Evaluations of Data Mining Methods in Order to Provide the Optimum Method for Customer Churn Prediction: Case Study Insurance Industry."2012 International Conference on Information and Computer Applications (ICICA 2012) IPCSI vol. 24 (2012) © (2012) IACSIT Press, Singapore.

[8]. Surjeet Kumar Yadav and Saurabh Pal:" Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification "World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 2, No. 2, 51-56, 2012.

[9]. K. Wisaeng "A Comparison of Different Classification Techniques for Bank Direct Marketing"International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-3, Issue-4, September 2013