International Journal of Engineering and Technical Research (IJETR)
ISSN: 2321-0869 (O) 2454-4698 (P), Volume-6, Issue-4, December 2016

Performance Analysis of Text Classification Algorithms using Confusion Matrix

Maria Navin J R, Pankaja R

***Abstract*— Text classification under text mining plays an important role in the process of classifying digital documents. Proper classification of documents needs techniques like text mining, natural language processing and machine learning to obtain meaningful knowledge. In this paper we have presented performance analysis of text classification algorithms namely k-Nearest Neighbor, Naïve Bayes, logistic regression and Support Vector Machines by creating confusion matrices for training and testing sets obtained by applying 10-fold cross validation method on a corpus of movie review data set. A comparative study is performed on the performance of the algorithms by computing statistical parameters like accuracy, kappa, sensitivity, specificity, positive and negative prediction value, prevalence, detection rate, detection prevalence and balanced accuracy from the confusion matrices.**

***Index Terms*— document classification, confusion matrix, support vector machine, movie reviews, corpus, accuracy, cross validation.**

I. INTRODUCTION

In recent research text classification and related concepts have become important due to the increasing number of sources that generate electronic documents according to [1] different text mining techniques are required to analyse data on social networking websites to identify various textual patterns. In text classification we classify the documents based on predefined categories. A corpus can be used as a main structure for managing and representing a collection of text documents. Preprocessing of text data like removing punctuations, numbers, converting text to lower case, eliminating synonyms and stemming is performed on the corpus. The performance of a document classification technique can be obtained by creating confusion matrices on training and testing data sets.

k-Nearest Neighbor, Naïve Bayes, Logistic Regression and Support Vector Machines are the key document classification techniques. In k-Nearest Neighbor the degree of similarity among documents is checked to obtain its neighbors and thereby the category of the text document can be found. This technique is known for its simplicity and its performance on classification on multiple classes. But the major drawback of k-Nearest Neighbor is that its performance decreases due to high presence of noise [2].

Naïve Bayes classifier performs prediction using conditional probability. In this method conditional probability of different words in the same categories are independent. The basic idea in Naïve Bayes approach is to use the joint probabilities of words and categories to estimate the probabilities of categories given a document.

Logistic Regression analysis is a binary prediction probabilistic statistical model for predicting between classes of a dataset by constructing a hyper plane. The class membership is obtained by probability measure that includes regression coefficients, intercepts and risk factors [3].

Support Vector Machines is the most commonly used algorithm for document classification. The key aspect of Support Vector Machines is to construct a hyper plane between the classes that provides maximum margins and use these boundaries for text classification. For a two dimensional case the created hyper plane is a straight line. Main advantage of Support Vector Machines is that it can datasets with many attributes with less over fitting than other methods. However, SVM classification has disadvantages in limited speed during both training and testing phases [4].

According to [5] Plug load identification using different machine learning algorithms were tested on a dataset. The study shows that k-Nearest Neighbor yields good accuracy in identifying the device. In [6] four supervised learning assessments were evaluated using 10-fold cross validation, the performance of k-Nearest Neighbor algorithm was stable across test options and performs best.

Decision Tree Classifier performs than Naïve Bayes and k-Nearest Neighbor better in terms of classification time [7]. Based on Precision, Recall, F-measure, Accuracy Naïve Bayes performs better. Except precision Naïve Bayes outperforms Decision Tree and k-Nearest Neighbor on all parameters.

Decision Tree has good applicability for diagnosis and prognosis and shows good performances. However the performance of Logistic Regression and Random forest classifier depends on the ratio of discriminators and performs better with higher number of discriminators [8].

II. METHODOLOGY

A. Movie Review Data Set

Binary classification is one of the most popular classification task [9]. The movie reviews corpus which is a collection of both positive and negative reviews can be used for opinion and sentiment-analysis experiments. The movie reviews documents in the corpus are labeled with respect to their overall sentiment polarity either positive or negative. This is a collection of movie reviews used for various opinion analysis tasks for predicting polarity ratings. This dataset is split into 1000 positive and 1000 negative reviews as pdf documents. These binary movie reviews can be classified

Maria Navin J R, Asst. Prof., Dept. of ISE, Sri Venkateshwara College of Engineering, Bangalore, India, +919738525948
Pankaja R, Asst. Prof., Dept. of ISE, Sri Venkateshwara College of Engineering, Bangalore, India, +919880251249

using k-Nearest Neighbor, Naïve Bayes, Logistic Regression and Support Vector Machines document classification techniques.

B. *Cross Validation*

Cross validation is a technique used to assess and evaluate the performance of machine learning algorithms. This technique is applied on new datasets that is not yet trained. This can be done by partitioning a dataset into two subsets one for training and the other for testing. In every round of cross validation we randomly partition the given original data set into a training set that is used for training a machine learning algorithm and testing set for evaluating its performance.

In our experiments, we have used a 10-fold cross validation, where the dataset is partitioned into 10 equal partitions and each is used for validation and the remainder is used for training in all possibilities.

C. *Confusion Matrix*

The evaluation of document classification techniques can be obtained in terms of correctness by computing statistical measures namely the True Positives (TP), True Negatives (TN), False Positive (FP) and False Negatives (FN). These components form the Confusion Matrix as shown in Fig. 1.

A confusion matrix is a table that can be generated for a classifier on a binary data set and can be used to describe the performance of the classifier.

	Predicted: No	**Predicted: Yes**
Actual: No	TN	FP
Actual: Yes	FN	TP

Fig. 1 Confusion Matrix

This matrix is based on the terms
T**rue Positives (TP) -** prediction and actual both are yes.
True Negatives (TN) - prediction is no and actual is yes.
False Positives (FP) - prediction is yes and actual is no.
False Negatives (FN) - prediction is no and actual is no.

D. *Performance Measures using Confusion Matrix*

P**ositive predictive value** (PPV) is the ratio of true positives and overall positives calls.
N**egative predictive value** (NPV) is the ratio true negatives and overall negatives calls.
Kappa Score κ will be high if there is a big difference between the accuracy and the null error rate.
Accuracy describes how often classifier is correct.
True Positive Rate or **Sensitivity** is the ratio of True Positives to overall actual yes.
False Positive Rate is the ratio of False Positives to overall actual no.
Specificity is the ratio of True Positives to overall actual no.
Precision is the ratio of True Positives to overall predicted yes.
Prevalence is the ratio of actual yes to total number of instances.
Detection Rate is the rate of true happenings also predicted to be happenings.
Detection Prevalence is the prevalence of predicted events.
Balanced Accuracy is the average of sensitivity and specificity [10].

III. Experimental Results and Analysis

The performance study on movie reviews data set was carried using R 3.1.2 software which is freely available. The movie reviews dataset contains a total of 2000 sample documents. To remove the bias, a 10-fold cross validation method is used for creating the training and testing sets.

In this paper we have used tm library under R for creating corpora which holds positive and negative reviews. R library SnowballC is used for pre-processing and inspecting the corpus. Pre-processing like converting to lower case, removing numbers and punctuations, stemming and stripping of white spaces are done. From the corpora we create a term-document matrix which represents terms and documents relationship in the form of a matrix where each term is a row, each column is a document and an entry is the frequency of occurrences of the term in the document. We use the function DocumentTermMatrix() in R to create the matrix and this matrix can be used to obtain length of the reviews which may be related to either positive or negative. This length attribute plays an important role in this classification.

Performance analysis of Document classification algorithms using 10-fold cross validation on total of 2000 movie reviews (1000 positives and 1000 negative reviews) for training and testing set. Cross validation is used for randomly choosing training and testing data sets in the ratio of 9/10 (N=1800) for training set and 1/10 (N=200) testing set respectively.

k-Nearest Neighbor algorithm is implemented in R using the class library. The Confusion Matrix is obtained by using confusionMatrix() function from caret library [11] and applied on a sample of 10-fold cross validated training and testing sets. The Confusion Matrices of the training and testing sets of k-Nearest Neighbor, Naïve Bayes, Logistic Regression and Support Vector Machines are shown from Table 1 to Table 8.

N=1800	**Predicted: No**	**Predicted: Yes**
Actual: No	698	214
Actual: Yes	206	682

Table 1: Confusion Matrix for the Training Set using k-Nearest Neighbor

N=200	**Predicted: No**	**Predicted: Yes**
Actual: No	79	20
Actual: Yes	22	79

Table 2: Confusion Matrix for the Testing Set using k-Nearest Neighbor

Naïve Bayes is implememented using in R using the e1071 library. The Confusion Matrix is obtained on a sample of 10-fold cross validated training and testing sets.

N=1800	**Predicted: No**	**Predicted: Yes**
Actual: No	629	275
Actual: Yes	176	720

Table 3: Confusion Matrix for the Training Set using Naïve Bayes

N=200	**Predicted: No**	**Predicted: Yes**
Actual: No	78	23
Actual: Yes	22	77

Table 4: Confusion Matrix for the Testing Set using Naïve Bayes

Logistic Regression is implemented in R using the glm() function. The Confusion Matrix obtained on a sample 10-fold cross validated training and testing sets using Logistic Regression.

N=1800	**Predicted: No**	**Predicted: Yes**
Actual: No	736	168
Actual: Yes	161	735

Table 5: Confusion Matrix for the Training Set using Logistic Regression

N=200	**Predicted: No**	**Predicted: Yes**
Actual: No	87	13
Actual: Yes	11	89

Table 6: Confusion Matrix for the Testing Set using Logistic Regression

Support Vector Machine is implemented in R using the library e1071. The Confusion Matrix obtained on a sample 10-fold cross validated training and testing sets using Support Vector Machine.

N=1800	**Predicted: No**	**Predicted: Yes**
Actual: No	817	87
Actual: Yes	84	812

Table 7: Confusion Matrix for the Training Set using Support Vector Machine

N=200	**Predicted: No**	**Predicted: Yes**
Actual: No	95	6
Actual: Yes	6	93

Table 8: Confusion Matrix for the Testing Set using Support Vector Machine

Comparison of Statistical Parameters for k-Nearest Neighbor, Naïve Bayes, Logistic Regression and Support Vector Machine obtained from respective Confusion Matrices of training and testing sets are shown in Table 9 and Table 10.

Performance Measures	**k-Nearest Neighbor**		**Naïve Bayes**	
	Training Set	**Testing Set**	**Training Set**	**Testing Set**
Accuracy	0.7667	0.79	0.7494	0.775
Kappa	0.5333	0.72	0.4991	0.55
Sensitivity	0.7721	0.7821	0.7814	0.78
Specificity	0.7612	0.7979	0.7236	0.77
Pos Pred Value	0.7654	0.7979	0.6958	0.7722
Neg Pred Value	0.7680	0.7821	0.8036	0.7777
Prevalence	0.5022	0.505	0.4472	0.5
Detection Rate	0.3878	0.395	0.3494	0.39
Detection Prevalence	0.5067	0.5008	0.5022	0.505
Balanced Accuracy	0.7666	0.7900	0.7525	0.775

Table 9: Performance Measures using k-Nearest Neighbor and Naïve Bayes algorithms

Performance Measures	**Logistic Regression**		**Support Vector Machine**	
	Training Set	**Testing Set**	**Training Set**	**Testing Set**
Accuracy	0.8172	0.88	0.905	0.94
Kappa	0.6344	0.7	0.81	0.878
Sensitivity	0.8205	0.8877	0.9068	0.9405
Specificity	0.8140	0.8725	0.9032	0.9393
Pos Pred Value	0.8142	0.87	0.9038	0.9405
Neg Pred Value	0.8203	0.89	0.9063	0.9393
Prevalence	0.4983	0.499	0.5006	0.505
Detection Rate	0.4089	0.435	0.4539	0.475
Detection Prevalence	0.5022	0.5	0.5022	0.505
Balanced Accuracy	0.8172	0.8801	0.9050	0.9399

Table 10: Performance Measures using Logistic Regression and Support Vector Machine

IV. CONCLUSION

A comparative analysis on the performance of k-Nearest Neighbor, Naïve Bayes, Logistic Regression and Support Vector Machines document classification techniques is presented in this paper. R 3.1.2 statistical package was used to obtain the results and the implementation of the document classification techniques were done using the R libraries tm, class, caret, e1071, SnowballC and R functions DocumentTermMatrix() and glm(). Confusion Matrices were created for training and testing data sets using 10-fold cross validation method on the corpora with movie reviews data sets with 2000 reviews. Statistical measures namely accuracy, kappa, sensitivity, specificity, positive predictive value, negative predictive value, detection rate, detection prevalence and balanced accuracy were computed using the confusion matrices. We conclude that the statistical parameters calculated from the respective confusion matrices

indicated that support vector machine performs better classification than Logistic Regression, k-Nearest Neighbor and Naïve Bayes techniques.

REFERENCES

[1] Rizwana Irfan et al, "*A survey on text mining in social network*" in The Knowledge Engineering Review, Vol. 30:2, Cambridge University Press, 2015, pp. 157–170.

[2] Aurangzeb Khan et al, "*A Review of Machine Learning Algorithms for Text-Documents Classification*", Journal of Advances in Information Technology, vol. 1, no. 1, February 2010.

[3] Kanthida Kusonmano et al, "*Evaluation of the Impact of Dataset Characteristics for Classification Problems in Biological Applications*", World Academy of Science, Engineering and Technology International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering Vol:3, No:10, 2009.

[4] Jinho Kim, Byung-Soo Kim, Silvio Savarese, "*Comparing Image Classification methods Nearest Neighbor, Support Vector Machines*", Applied Mathematics in Electrical and Computer Engineering, ISBN: 978-1-61804-064-0.

[5] Raghunath Shivram Reddy, Niranjan Keesara, Vikram Pudi and Vishal Garg, "*Plug load identification in educational buildings using machine learning algorithms*", Proceedings of BS2015:14th Conference of International Building Performance Simulation Association, Hyderabad, India, Dec. 7-9, 2015.

[6] Olufemi Sunday Adeoye, "*The Effect of the Different Supervised LearningAssessments on the Performance of Machine Learning Schemes*", International Journal for Research In Emerging Science And Technology, Volume-3, Issue-4, Apr-2016.

[7] Ahmad Ashari, Iman Paryudi, A Min Tjoa, " *Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool",* International Journal of Advanced Computer Science and Applications, Vol. 4, No. 11, 2013.

[8] Kanthida Kusonmano, Michael Netzer, Bernhard Pfeifer, Christian Baumgartner, Klaus R. Liedl and Armin Graber *"Evaluation of the Impact of Dataset characteristics for Classification Problems in Biological Applications",* World Academy of Science, Engineering and Technology International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering Vol:3, No:10, 2009.

[9] Marina Sokolova, Guy Lapalme, *"A systematic analysis of performance measures for classification tasks",* Information Processing and Management, 45 (2009), Elsevier Ltd. 427–437.

[10] Vapnik, V. (1998) Statistical Learning Theory. John Wiley & Sons.

[11] Chichester.Kuhn, M. (2008), "Building predictive models in R using the caret package" Journal of Statistical Software.