Comparison of Text Classifiers on News Articles

Lilima Pradhan[1], Neha Ayushi Taneja[2], Charu Dixit[3] ,Monika Suhag[4]

[1]*Lilima Pradhan, Department of Computer Engineering, Army Institute of Technology, Pune, Maharashtra*
[2]*Neha Ayushi Taneja, Department of Computer Engineering, Army Institute of Technology, Pune, Maharashtra*
[3]*Charu Dixit, Department of Computer Engineering, Army Institute of Technology, Pune, Maharashtra*
[4]*Monika Suhag, Department of Computer Engineering, Army Institute of Technology, Pune, Maharashtra*

---***---

Abstract - *Text classification is used to classify documents on the basis of their content. The documents are assigned to one or more categories manually or with the help of classifying algorithms. There are various classifying algorithms available and all of them vary in efficiency and the speed with which they classify documents. The news articles classification was done using Support Vector Machine(SVM) classifier, Naive Bayes classifier, Decision Tree classifier, K-nearest neighbor(kNN) classifier and Rocchio classifier. It is found that SVM gives a higher accuracy in comparison with the other classifiers tested.*

***Key Words*: Text Classification, Support Vector Machine, Naive Bayes, k-Nearest Neighbor, Decision Tree, Rocchio, Preprocessing**

1.INTRODUCTION

Classification is the task of choosing the correct class label for a given input. In basic classification tasks, each input is considered in isolation from all other inputs, and the set of labels is defined in advance [1]. With the rapid growth of technology, the amount of digital information generated is enormous and so organizing the information is very important. Classifying the news articles according to their content is highly desirable as it can enable automatic tagging of articles for online news repositories and the aggregation of news sources by topic (e.g. google news), as well as provide the basis for news recommendation systems [2]. Text classification of news articles helps in uncovering patterns in the articles and also provide better insight into the content of the articles. The aim of the paper is to present a comparison of the various popular classifiers on various datasets to study the characteristics of the classifiers[3][4].

2. CLASSIFIERS

2.1 Naive Bayes

Naive Bayes is an classification algorithm based on Bayes Theorem. The Bayes Theorem is used for finding conditional probabilities and conditional probability is used to denote the likelihood of occurrence of an event given an event which has previously occurred i.e. it uses the knowledge of prior events to predict the future events. Using Bayes Theorem, the conditional probability can be decomposed as [5]:

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}} \qquad \text{... (1)}$$

In the context of text classification, the probability that a document d_j belongs to a class c is calculated by the Bayes theorem as follows [6]:

$$p(c|d_j) = \frac{p(d_j|c)p(c)}{p(d_j)} = \frac{p(d_j|c)p(c)}{p(d_j|c)p(c) + p(d_j|\bar{c})p(\bar{c})}$$
$$= \frac{\frac{p(d_j|c)}{p(d_j|\bar{c})} \cdot p(c)}{\frac{p(d_j|c)}{p(d_j|\bar{c})} \cdot p(c) + p(\bar{c})} \qquad \text{... (2)}$$

It makes an assumption that all the features are independent. Inspite of the assumption, Naive Bayes works quite well for real world problems like spam filtering and text classification. Its main advantage is that it requires small training data to estimate necessary parameters.

1.2 Decision Tree

Decision tree is a very popular tool for classification and prediction. It represents rules that are used to classify data. Decision tree has tree like structure where leaf node indicates class and intermediate node represents decision. It starts with a root node and then branches into multiple solutions. An attribute or branch is selected using different measures like gini index, information gain and gain ratio. Its aim is to predict the value of a target variable based on different inputs by learning simple decision rules. Decision tree uses recursive approach which divides source set into subsets based on attribute value test. Many algorithms can be used for building decision tree like ID3, C4.5, CARD(Classification and Regression Tree) and CHAID.

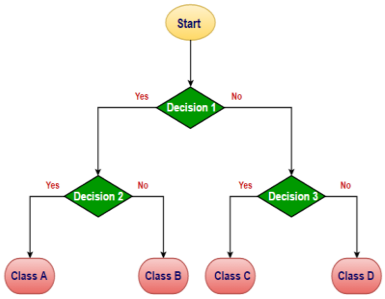

Fig-1: Decision Tree Classifier

1.3 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm mainly used in classification and regression. In this, we plot each data item as a point in n-dimensional space(n is number of features). Then for classification we find a hyper-plane which best divides the different classes(maximizes the distance between the data and the nearest data point in each class). This hyper-plane can be a straight line(linear) or any curve. SVM uses Kernel trick for performing non-linear classification. Kernel is used for transforming low dimensional space into high dimensional space. There are three types of kernel i.e. linear, polynomial and radial. SVM can be used in many real world problems like image classification, text classification, hand-written character recognition, etc.

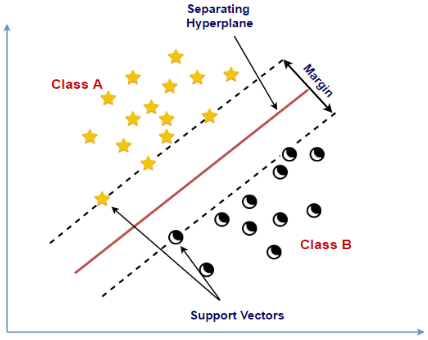

Fig-2: Hyperplane separating two classes

1.4 k-Nearest Neighbor

k-Nearest Neighbor(kNN) is one of the simplest machine learning algorithm used for regression and classification. It is a lazy algorithm that means there is no explicit training phase or it is very less. In this classifier, data is represented in a vector space. For unlabelled data, kNN looks for k points which are nearest to that unlabelled data by selecting distance measure. These K points becomes the nearest neighbours of that unlabelled data. Unlabelled data is assigned the class which represents the most objects among those k neighbours. Mainly three distance functions are used for finding k nearest neighbours, these are : euclidean distance, manhattan distance and minkowski distance. For finding optimal k value, first segregate the training and validation dataset and then plot the validation error curve. It is a versatile algorithm and is applied in large number of fields.

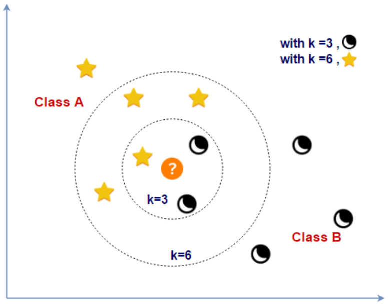

Fig-3: kNN Classifier

1.5 Rocchio

It is based on relevance feedback method. To increase the recall and precision as well, documents are ranked as relevant and non-relevant. Nearest centroid is also known as Rocchio classifier when it is applied to text classification using tfidf vectors. In nearest centroid, data is assigned to the class whose centroid is nearest to the data.

2. PREPROCESSING OF DATA

The news articles are first read from the files in raw form and then various types of processing is done on the text data. The dataset is first split into training and testing data(the split that gives the maximum efficiency for the classifiers is chosen, keeping in mind that the model is not overfitted). Then the cleaning of data is done by removing non-unicode characters, removing stop words and stemming [7]. The data is then converted to vector form and tf-idf is applied to find the importance of a word to the document. Feature selection techniques is applied to select the best features by applying techniques like Chi-Square test. Then the classifiers are trained on the train data and then tested for obtaining the accuracy, training time and testing time for each dataset.

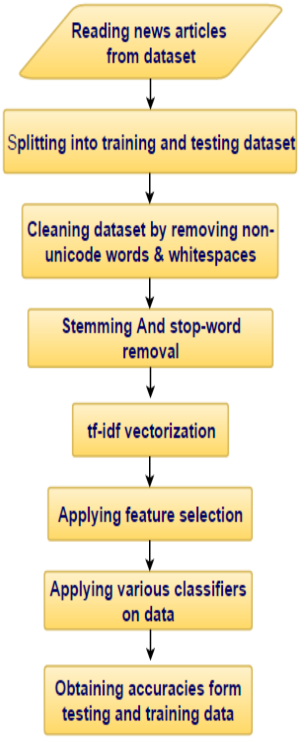
Reading news articles from dataset
Splitting into training and testing dataset
Cleaning dataset by removing non-unicode words & whitespaces
Stemming And stop-word removal
tf-idf vectorization
Applying feature selection
Applying various classifiers on data
Obtaining accuracies form testing and training data

Fig-4: System Design for Classification of news articles

3. EXPERIMENTAL RESULTS AND OBSERVATIONS

The news articles datasets used for the comparison of classifiers are Reuters dataset[8], Twenty newsgroups dataset[9], BBC News dataset, BBCSport News dataset[10]. The Twenty newsgroups dataset is collection of approximately 20000 articles under 20 categories and the articles are distributed uniformly, each article stored as a separate file. The Reuters dataset is a collection of documents that appeared on Reuters newswire in 1987. The documents were assembled and indexed with categories. The two BBC news article datasets, originating from BBC News, are provided for use as benchmarks for machine learning research. These datasets are made available for non-commercial and research purposes only. The five classifiers are applied on each of the above datasets and results are obtained for all the datasets. Some datasets have uniform document distribution while others are non-uniform.

The text for classification is of varying lengths and thus needs to be processed properly for accurate and precise results. Text cleaning, stemming, stop word removal and other preprocessing is performed as explained above. Three fourth of the dataset has been taken as training set and the rest as testing set(the data split ratio has been chosen by performing efficiency analysis on different splits of data). It is important to find the correct train to test ratio as it effects the training of classifier to predict correct categories for test set.

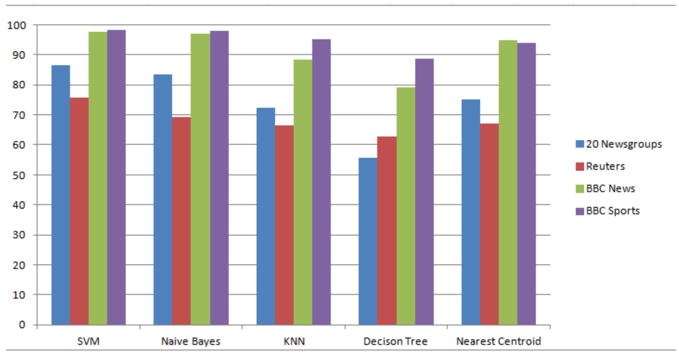

Fig-5: Accuracy of different classifiers on news datasets

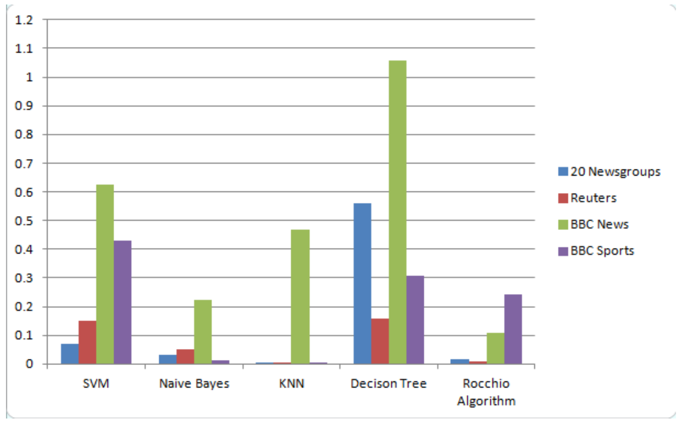

Fig-6: Training time of different classifiers on news articles datasets

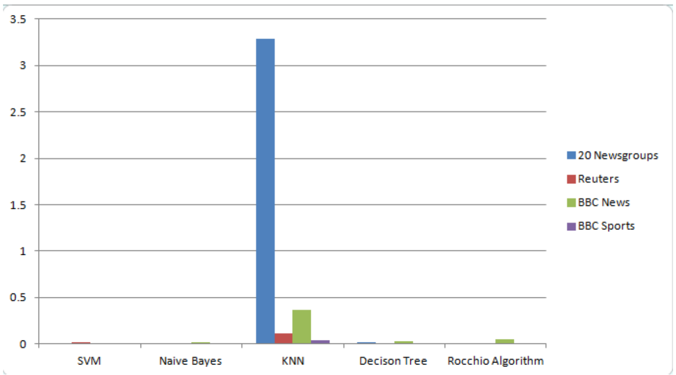

Fig-7: Testing Time of different classifiers on news articles datasets

Table -1: Accuracy of different classifiers

ACCURACY	
Twenty Newsgroups	
SVM	86.7097
Naive Bayes	83.5779
K-Nearest Neighbors	72.3429
Decision Tree	55.6664
Rocchio Algorithm	75.0141
Reuters	
SVM	75.6709
Naive Bayes	69.2594
K-Nearest Neighbors	66.5258
Decision Tree	62.6988
Rocchio Algorithm	67.2216
BBC News	
SVM	97.6744
Naive Bayes	96.9588
K-Nearest Neighbors	88.5509
Decision Tree	79.0697
Rocchio Algorithm	94.8121
BBC Sports	
SVM	98.3870
Naive Bayes	97.8494
K-Nearest Neighbors	95.1612
Decision Tree	88.7096
Rocchio Algorithm	94.0860

Table -2: Testing and training time of different classifiers

	TRAINING (minutes)	TESTING (minutes)
Twenty Newsgroups		
SVM	0.0708054	0.00072495
Naive Bayes	0.0327356	0.00815615
K-Nearest Neighbors	0.0054167	3.28215916
Decision Tree	0.5594385	0.01354642
Rocchio Algorithm	0.0148767	0.00490734
Reuters		
SVM	0.1493567	0.01644477
Naive Bayes	0.0494006	0.00391138
K-Nearest Neighbors	0.0031509	0.11639163
Decision Tree	0.1587691	0.00006292
Rocchio Algorithm	0.0072536	0.00088035
BBC News		
SVM	0.6247513	0.00202059
Naive Bayes	0.2239165	0.01982212
K-Nearest Neighbors	0.4692099	0.36869502
Decision Tree	1.0578448	0.02949571
Rocchio Algorithm	0.1088447	0.05044627
BBC Sports		
SVM	0.4297354	0.00075793
Naive Bayes	0.0123956	0.00753688
K-Nearest Neighbors	0.0032143	0.03960132
Decision Tree	0.3084764	0.00121045
Rocchio Algorithm	0.2423000	0.00314807

For each dataset a different level of accuracy was obtained. In Twenty newsgroups dataset the data is divided uniformly among the categories. The best accuracy and f-score was observed for Linear SVM with accuracy of 86.7%.The required training and testing time was also small. Similarly for Reuters Dataset the best accuracy was obtained for Linear SVM with 75.67% accuracy. The Reuters dataset is large but not as uniform as Twenty newsgroups ,due to which accuracy values were not that high.

Two other datasets BBC News and BBC Sports were also used. These datasets were smaller in comparison to Newsgroups 20 and Reuters. Linear SVM gave the best accuracy in both datasets.

4. CONCLUSION

The comparison of the classifiers on the four datasets yields the result that linear SVM gives the highest accuracy for classifying the news articles followed by Naive Bayes.[11]. The Decision Tree classifier takes the most time to train itself using the training data followed by SVM, although both take a few seconds to train as well as test the data. K-nearest neighbour takes the most time to test data as it performs minimal computation during testing phase and does most of the computation while testing , and also has high memory requirements as training data needs to be in memory to correctly find labels for new data. The experimentation also reveals that the classifying algorithms perform better for smaller datasets, but to support the obervation the classifiers need to be tested for many other datasets of varying sizes. The experiments show that linear SVM should be preferred for better accuracy while Naive Bayes is a good alternative with better time complexity although a little less accuracy.

3. FUTURE WORK

The classification of the news articles could be expanded to multi-label text articles, as there are many news articles that do not belong to a single category and are instead a mixture of various categories. Being able to classify multi-label articles will provide added advantage to the classifiers.
Also systems like news articles recommender can be implemented to test the efficiency of the underlying classifier algorithms and also help users to enjoy a hassle free and better reading experience.

ACKNOWLEDGEMENT

We would sincerely like to thank our project guide Prof Yogita T Hambir for her invaluable guidance all through the project work.

REFERENCES

[1] Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc.

[2] Chase, Zach, Nicolas Genain, and Orren Karniol-Tambour. "Learning Multi-Label Topic Classification of News Articles."

[3] Wang, Yaguang, et al. "Comparison of Four Text Classifiers on Movie Reviews." Applied Computing and Information Technology/2nd International Conference on Computational Science and Intelligence (ACIT-CSI), 2015 3rd International Conference on. IEEE, 2015.

[4] Ramdass, Dennis, and Shreyes Seshasai. "Document classification for newspaper articles." (2009).

[5] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[6] Kim, Sang-Bum, et al. "Some effective techniques for naive bayes text classification." IEEE transactions on knowledge and data engineering 18.11 (2006): 1457-1466.

[7] Albishre, Khaled, Mubarak Albathan, and Yuefeng Li. "Effective 20 Newsgroups Dataset Cleaning." Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015 IEEE/WIC/ACM International Conference on. Vol. 3. IEEE, 2015.

[8] http://disi.unitn.it/moschitti/corpora.html

[9] https://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups

[10] http://mlg.ucd.ie/datasets/bbc.html

[11] Dilrukshi, Inoshika, and Kasun De Zoysa. "Twitter news classification: Theoretical and practical comparison of SVM against Naive Bayes algorithms." Advances in ICT for Emerging Regions (ICTer), 2013 International Conference on. IEEE, 2013.