IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 74-78
www.iosrjournals.org

A Comparative Study of Machine Learning Approaches for Text Classification

Lino Murali[1], Anjali A[2], Athira Raj[3]
Department of Computer Science and Engineering Sree Narayana Gurukulam College of Engineering Kadayiruppu, India.

***Abstract*:** Perhaps the single largest data source in the world is the world wide web. Heterogeneous and unstructured nature of the data on web has challenged mining the web. Practical needs to extract textual information and unseen patterns continue to drive the research interest in text mining. Faultless categorization of texts can be better performed by machine learning techniques. In this paper we present a review of various text classification approaches under machine learning paradigm. Existing classification algorithms including Decision Tree, Naive Bayes, Support Vector Machine and k-Nearest Neighbors are compared based on speed, accuracy, interpretability and multi-class support.

Keywords: *Text classification; Machine Learning; Naive Bayes; k-Nearest Neighbour; Support Vector Machine; Decision Tree.*

I. Introduction

The World Wide Web serves as a huge repository of highly dynamic and diverse data that is growing at an exponential rate. This available information must be organized systematically for its proper utilization. Systematic organization of information facilitates retrieval of relevant text content for needy application including information search and retrieval, personalization etc [6]. Text mining can help in organizing text-based content and derive valuable insight from textual data. A key aspect in text mining includes the classification of the documents. To gather useful information from these, the text has to be categorized. The task to classify a given data instance into a pre-specified set of categories is known as Text Classification. It is highly important to develop techniques to classify text documents accurately and efficiently.

II. Text Classification

Text classification refers to the process of classifying the documents based on their content into a certain number of pre-defined classes. Each document can be in no, multiple or exactly one category. The goal of Text Classification is to assign a category to a new document[2].

Text Classification can be done either manually or automatically.Performing the task manually can be expensive and time consuming. Automatic Text Classification is an effective classification method to deal with voluminous data over the internet. Automatic classification of text is an important problem in many domains. For instance, text classification finds its application in automatic spam detection, sentiment analysis, indexing of scientific articles, personal email sorting , classification of news articles etc.

During the recent years, the text categorization technique using machine learning techniques has observed an active attention. This paper illustrates the text classification process using machine learning techniques. Various machine learning approaches have been discussed which includes Naive Bayes(NB), k-Nearest Neighbour(k-NN), Decision Tree(DT) and Support Vector Machine(SVM).

III. Machine Learning Classifiers

Machine learning is concerned with the design and development of algorithms and techniques that allow computers to "learn" so as to improve the expected future performance[6]. Machine learning based automatic text classification is a supervised learning method. A supervised learning approach takes a known set of input data and known output for the data and trains a model to generate predictions for the response to new data.

A wide range of supervised machine learning approaches are available and there is no single approach that works best on all classification problems. The purpose of this study is to review the available and known machine learning approaches for text classification and comparative study of these different methods. The approaches discussed in this paper includes Decision Tree, Naive Bayes , k-NN, Support Vector Machine.

A. Decision Tree

A decision tree classifier is a tree structure whose features can be exploited for text classification. A document to be tested can be given to the root and the tree classifies the document by recursively testing it until it reaches a leaf node. Each leaf node is labeled by categories and this category can be assigned to the document. Every internal nodes are labeled and they check the incoming document based on these labels and forward to the proper sub-tree. Finally the label of leaf node where the document reached represents its category. Decision tree is experimentally simple and understandable but classifier performance degrades when dealing with continuous attributes.

B. Naive Bayes

Naive Bayes classifiers are simple probabilistic classifiers based on Bayes' theorem with strong independence assumptions between the features [8]. The basic idea is to find the probability that the document belongs to the class with the highest posterior probability. The Bayes theorem is used to estimate the posterior probability. Given a problem instance to be classified, represented by a vector d = (d1,d2,...,dn) representing some n features, for each of K possible classes , it assigns to this instance probabilities

$$p(Ck| d1,d2,...,dn)$$

Using Bayes' theorem, the conditional probability can be calculated as

$$p(Ck|d)=\underline{p(Ck)p(d|Ck)}\ p(d)$$

p(Ck|d)is the posterior probability of class p(Ck) is the prior probability of class

p(d|Ck)is the likelihood which is the probability of predictor given class
p(d) is the prior probability of predictor

The Bayes classifier assumes variable independence, only the variances of the variables for each class need to be determined and the presence of one feature does not affect any other features in classification process. This makes Bayes classifiers simple to implement and work well on numeric and textual data. Bayes classifiers can be trained very efficiently by requiring a relatively small amount of training data to estimate the parameters necessary for classification.

The classifier has a relatively low classification performance when features are highly correlated. When compared to other discriminative algorithms, the Naive Bayes classification approach has relatively low classification performance. In case of text classification, the classifier ignores the frequency of word occurrences in the feature vector.

C. k-Nearest Neighbour

The k-NN algorithm is widely used in text classifications due to its effectiveness, non-parametric and easy to implement properties. k-NN is an instant-based learning algorithm that classifies objects based on closest feature in the training set. An instance is categorized by a majority of its neighbours, with the instance being assigned to the class most common among its k nearest neighbors. The neighbours are taken from the training set of instances for which the class is known.

Usually Euclidean distance is typically used in computing the distance between the vectors. In case of text classification, another metric can be used, such as the Hamming distance. The key aspect of this method is the availability of members of the same class with similar characteristics.

The performance of a k-NN classifier primarily depends on the optimal value of k chosen as well as the distance metric applied. k-NN method performs well even in handling the classification tasks with multi-categorized documents[1]. A major limitation of k-NN is that it uses all features in computing distances. With respect to text classification, only smaller number of the total vocabulary may be useful in categorizing documents. Other limitations include longer classification time and higher computation cost.

D. Support Vector Machine

Support Vector Machine (SVM) is a supervised classification technique from the machine learning

field that has been successfully applied for text classification task[1].

SVM use the principle of Structural Risk Minimization (SRM) from computational learning theory. SRM can guarantee the lowest true error by finding a hypothesis h. The true error of h is find by taking the probability that h makes an error on a new test sample selected randomly. There is an upper bound to connect the true error of a hypothesis h and the error of h on the training set. Thus it finds all the surfaces in n-dimensional space which separate the positives from the negatives training samples by the widest possible margin[4].

The aim of SVM is to find a hyper plane that clearly separates the positive and negative data points in an n-dimensional space. Among many such hyper planes in the high dimensional space SVM chooses one hyper plane which has maximum margin distance to the nearest data points. These points are called support vectors.

Margin can be calculated by constructing two parallel hyper planes, one on each side of the separating hyper-plane, which are "pushed up against" the two data sets. Intuitively, a good separation is achieved by the hyper-plane that has the largest distance to the neighboring data points of both classes, since in general the larger the margin the lower the generalization error of the classifier[8].

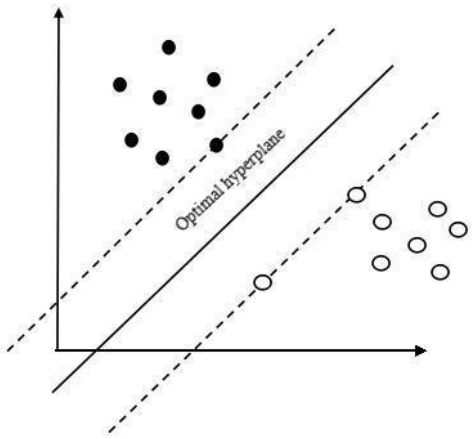

Figure 1. Optimal hyper-plane

IV. Comparative Observations

Several algorithms or combination of algorithms as hybrid approaches exists for the automatic classification of documents. Among these algorithms SVM, Naive Bayes, k-NN and Decision Tree are compared in this paper.

Decision Tree classifier which performs well with smaller data set classification have trouble dealing with noise in training data. Naive Bayes works well on textual data when compared with other algorithms, however conditional independence assumption is violated by real-world data and perform very poorly when features are highly correlated and does not consider frequency of word occurrences[8]. k-NN classifier continues to achieve very good results and scales up well with the number of documents but classification time is longer and is difficult to find optimal value of k. SVM is one of the most effective text classification methods as it is able to manage large spaces of features and high generalization ability, but this makes SVM algorithm relatively more complex which in turn demands high time and memory usage during training stage and classification stage[6]. SVM capture the inherent characteristics of the data better but it is applicable to only binary classification.

TABLE I. CHARACTERISTICS OF CLASSIFIERS

Classi-fier	Description	Inter-pretability	Predict-ion Speed	Avg predictive accuracy	Multi-class Support

DT	Non-parametric supervised learning method Good for a small set of features.	Easy	Fast	Lower	Yes
NB	Probabilistic classifier. Good parameter estimation with small number of training data.	Easy	Fast	Lower	Yes
k-NN	Instance based learning model Time consuming when number of training examples are large.	Hard	Medium	Medium	Yes
SVM	Discriminative classifier. Can handle large feature set.	Easy	Medium	High	No

As per the comparative study, the prediction speed of Decision Tree and Naive Bayes is found to outperform the other two algorithms. Even though SVM lags with respect to prediction speed, it gives a higher average predictive accuracy.k-NN shows a better predictive accuracy compared to Decision Tree and Naive Bayes specifically when the data set has continuous attributes. In terms of interpretability, Decision Tree, Naive Bayes, SVM surpasses k-NN. As SVM is well known for binary classification, it lacks multi-class support whereas the other three classifiers extends a multi-class support.

V. Conclusion

The classification of vast number of documents available in world wide web is a big challenge in text mining. Machine learning techniques can be efficiently applied in this area to classify unseen documents accurately. This paper provides a review of machine learning approaches including Naive Bayes, k-NN, Decision Tree and SVM for text classification. Each algorithm has its own advantages and disadvantages as described in section III. The existing classification methods are compared and contrasted based on various parameters namely prediction speed, average predictive accuracy, interpretability and multi-class support. After comparing the existing machine learning approaches for text classification it can be concluded that SVM classifier is weighted as one of the effective text classification method but limited to binary classification. Researches shows that no single representation scheme and classifier can be mentioned as a general model for any application and the selection of classifier depends on the nature of the problem.

References

[1] Iswarya, Radha V, "Ensemble learning approach in improved K Nearest Neighbor algorithm for Text categorization" ,IEEE International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 .
[2] Lilleberg J, Marshall,Yun Zhu, Yanqing Zhang,"Support vector machines and Word2vec for text classification with semantic features", IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCICC), 2015.
[3] Pratiksha Y. Pawar and S. H. Gawande, "A Comparative Study on Different Types of Approaches to Text Categorization", International Journal of Machine Learning and Computing vol. 2, no. 4, pp. 423-426, 2012.
[4] Holts A, Riquelme C, Alfaro R, "Automated Text Binary Classificationusing Machine Learning Approach" ,IEEE XXIX International Conference of the Chilean Computer Science Society (SCCC), 2010.
[5] Vandana KordeC, Namrata Mahender,"Textclassificationandclassifiers: ASURVEY",International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012.
[6] Shweta C. Dharmadhikari, Maya Ingle , Parag Kulkarni, "Empirical Studies on Machine Learning Based Text Classification Algorithms",An International Journal (ACIJ), Vol.2, No.6, November 2011.
[7] Gaurav S. Chavan, Sagar Manjare, Parikshit Hegde, Amruta Sankhe,"A Survey of Various Machine Learning Techniques for Text Classification",International Journal of Engineering Trends and Technology (IJETT) ,Volume 15 ,Sep 2014.

[8] Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee,"A Review of Machine Learning Algorithms for Text-Documents Classification", Journal Of Advances In Information Technology, February 2010.
[9] M. Ikonomakiss, Kotsiantisv, Tampakas ,"Text Classification Using Machine Learning Techniques",WSEAS Transactions On Computers, Issue 8, Volume 4, August 2005.