A Comparative Study on Different Types of Approaches to Text Categorization

Pratiksha Y. Pawar and S. H. Gawande, *Member, IACSIT*

***Abstract*—Text Categorization is a pattern classification task for text mining and necessary for efficient management of textual information systems. The documents can be classified by three ways unsupervised, supervised and semi supervised methods. Text categorization refers to the process of assign a category or some categories among predefined ones to each document, automatically. This paper presents a comparative study on different types of approaches to text categorization.**

***Index Terms*—Text categorizatin, classifier, documents.**

I. Introduction

Today huge amount of information are being associated with the web technology and the internet. To gather useful information from it these text has to be categorized. The task to classify a given data instance into a pre-specified set of categories is known as "*text categorization*" (TC). Given a set of categories (subjects, topics) and a collection of text documents, it is the process of finding the correct topic (or topics) for each document.

The expert's knowledge about the categories is directly used to categorize the documents. Most of the recent work on categorization is concentrated on approaches which require only a set of manually classified training instances that are much less costly to produce. A classifier is built by learning from a set of pre-classified examples. One of the drawbacks of supervised approaches is that they need to be trained on predefined positive and negative test samples or predefined categories. Efficiency of these models depends on the quality of the sample sets. With the enormous amount of data and different type of applications, it is not always possible to create these training sets or contextual categories manually.

TC may be formalized as the task of approximating the unknown target function Φ: D × C → {T, F} (that describes how documents ought to be classified, according to a supposedly authoritative expert) by means of a function called the classifier, where C = {c_1. . . c|C|} is a predefined set of categories and D is a (possibly infinite) set of documents. If Φ (d_j, c_i) = T, then d_j is called a positive example (or a member) of c_i, while if Φ (d_j, c_i) = F it is called a negative example of c_i.

There are two types of approaches to text categorization: rule based and machine learning based approaches. Rule based approaches mean ones where classification rules are defined manually and documents are classified based on rules. Machine learning approaches mean ones where classification rules or equations are defined automatically using sample labeled documents. This class of approaches has much higher recall but a slightly lower precision than rule based approaches. Therefore, machine learning based approaches are replacing rule based one for text categorization.

In this paper section II describes different types of text categorization, Comparative study and newly proposed approaches are explained in section III and IV. The Section V and Section VI state details of hybrid approaches and concluding remark respectively.

II. Different Types of Approaches

A. K. Nearest Neighbor

KNN is a classification algorithm as given in [1] where objects are classified by voting several labeled training examples with their smallest distance from each object. The k-nearest neighbor classification method is outstanding with its simplicity and is widely used techniques for text classification. This method performs well even in handling the classification tasks with multi-categorized documents.

Its disadvantage is that KNN requires more time for classifying objects when a large number of training examples are given. KNN should select some of them by computing the distance of each test objects with all of the training examples.

B. Rocchio's Algorithm

The algorithm in [2] is easy to implement, efficient in computation, fast learner and have relevance feedback mechanism but low classification accuracy.

C. Decision Trees

A Decision Tree text classifier in [3] is a tree in which internal nodes are labeled by terms, branches departing from them are labeled by the weight that the term has in the text document and leafs are labeled by categories. Decision Tree constructs using 'divide and conquer' strategy. Each node in a tree is associated with set of cases. This strategy checks whether all the training examples have the same label and if not then select a term partitioning from the pooled classes of documents that have same values for term and place each such class in a separate subtree.

D. Naïve Bayes Algorithm

Naïve Bayes classifier is a simple probabilistic classifier based on applying Baye's Theorem with strong independence assumptions. This algorithm computes the posterior probability of the document belongs to different classes and it assigns document to the class with the highest posterior probability. This probability model would be

Manuscript received may 9, 2012; revised June 13, 2012.

Authors are with the Department of Computer Engineering, Government College of Engineering & Research, Awasari, Pune, India (e-mail: patu_pawar@yahoo.co.in; shgawande@yahoo.co.in).

independent feature model so that the present of one feature does not affect other features in classification tasks [4].

E. Back propagation Network

In this method text is categorized by non-linear feed-forward neural network trained by Back propagation learning rule. That is we apply text for classifying the text under supervised learning.

There is strong reason for using ANS in text categorization. For the problems which cannot be solved sequentially or by sequential algorithms ANS provides the better solution. It is useful in recognizing complex patterns and performing nontrivial mapping functions.

F. Support Vector Machines (SVM)

A Support Vector Machine is a supervised classification algorithm that has been extensively and successfully used for text classification task.

High dimensional input space: When learning text classifiers, one has to deal with large number of features. Since SVM use over fitting protection, which does not necessarily depend on the number of features, they have the potential to handle these large feature spaces.

Most text categorization problems are linearly separable: All categories are linearly separable and so are many of the Reuters Tasks. The idea of SVMs is to find such linear separators.

III. Comparative Observations

If we compare decision trees and neural networks we can see that their advantages and drawbacks are almost complementary. For instance humans easily understand knowledge representation of decision trees, which is not the case for neural networks. Decision trees have trouble dealing with noise in training data, which is again not the case for neural networks, decision trees learn very fast and neural networks learn relatively slow, etc. decision tree learning is used to do qualitative analysis and neural learning is used to do subsequent quantitative analysis.

Naïve Bayes classifier is a very simple classifier which works very well on numerical and textual data. It is very easy to implement and computationally cheap when compared to any other classification algorithm. One of the major limitations of this classifier is that it performs very poorly when features are highly correlated. Also with respect to a text classification, it fails to consider the frequency of word occurrences in the feature vector. The main disadvantage of the Naive Bayes classification approach is its relatively low classification performance compare to other discriminative algorithms, such as the SVM with its outperformed classification effectiveness.

Nearest Neighbor classifier is very effective and it is non-parametric in nature. As compare to Rocchio algorithm more local characteristics of documents are considered. But the classification time is very long and finding the optimum value of k is difficult.

One advantage that SVM offer for TC is that dimensionality reduction is usually not needed, as SVMs tend to be fairly robust to over fitting and can scale up to considerable dimensionalities. SVM was initially applied to text categorization by Joachim's [10]. Joachim validated the classification performance of SVM in text categorization by comparing it with NB and KNN. Drucker adopted SVM for implementing a spam mail filtering system and compared it with NB in implementing the system. They showed that SVM was the better approach to spam mail filtering than NB. In spite of the advantage of SVM it has some limitations.

i] It is applicable to only binary classification. If a multiple classification problem is given, it should be decomposed into several binary classification problems using SVM.

ii] Problem in representing documents into numerical vectors, sparse distribution, since inner products of its input vector and training examples generates zero values very frequently.

If a suitable pre-processing is used with k-NN, this algorithm continues to achieve very good results and scales up well with the number of documents, which is not the case for SVM. As for Naive Bayes, it also achieved good performance.

As per the analysis support vector machine has more parameters than logistic regression and decision tree classifier, SVM has the highest classification precision most of the time, however SVM is very time consuming because of more parameters and requires more computation time. Compared to SVM, logistic regression is computationally efficient.

IV. Some Newly Proposed Approaches

In this section new approaches for text categorization are explained in details.

A. Neural Text Categorizer Acronyms

The approach in [14] proposes an alternative representation of documents to numerical vectors and a new supervised neural network as an approach to text categorization using the alternative representation in order to avoid the two problems: huge dimensionality and sparse distribution.

The advantage of the proposed neural network is that NTC can classify documents with its sufficient robustness with its smaller input size and iteration of learning than traditional approaches using numerical vectors. Therefore, NTC solves the first problem, huge dimensionality, completely. Since sparse distribution cannot exist in string vectors, the second problem is also addressed. Another advantage of NTC is that it provides transparency about its classification

B. Improving the Efficiency by Self-Organizing Map s

The proposed method of Hierarchical Self-Organizing Map [SOM] reduces the dimensionality of document vectors without essentially losing information contained in the full vocabulary. The advantages of this approach are scalability, topology representation, decreased computational time, improved categorization performance and meaningful information retrieval.

C. Soft-Supervised Learning

This approach proposes a new algorithm for graph-based SSL and use the task of text classification to demonstrate its benefits over the current state-of-the-art.

Text classification is multi-class problem. Training fully-supervised text classifiers requires large amounts of labeled data whose annotation can be expensive. As a result there has been interest in using SSL technique for text categorization.

V. Hybrid Approaches

A. Neural Networks Initialized with Decision Trees

This is a hybrid approach can be applied to the problem of text categorization and to test its performance relative to a number of other text categorization algorithms. This approach introduce the use of a hybrid decision tree and neural network technique to the problem of text categorization, because hybrid approaches decision tree learning is used to do qualitative analysis and neural learning is used to do subsequent quantitative analysis.

The proposed hybrid approach for text categorization task constructs the networks by directly mapping decision nodes or rules to the neural units and compresses the network by removing unimportant and redundant units and connections.

This method showed that hybrid decision tree and neural network approach improved accuracy in text classification task and are comparatively better than single decision tree or neural network initialized randomly text classifiers performance comparable to previous results.

B. Probabilistic Neural Network (PNN)

Recently, slightly modified versions of support vector machines, kNN and decision trees have been proposed to deal better with multi-label classification problems.

PNN [16] is a hybrid approach proposes a new version of a Probabilistic Neural Network (PNN) to tackle these kinds of problems. The proposed method compared against other classifiers. This approach is better than the other algorithms in many metrics typically well known in the literature for the multi-label categorization problems.

C. Bahes Formula for Classification

A New hybrid text document classification approach is proposed in [12], used naive Bayes method at the front end for raw text data vectorization, in conjunction with a SVM classifier at the back end to classify the documents to the right category. They shows that the proposed hybrid approach of the Naive Bayes vectorizer and SVM classifier has improved classification accuracy compared to the pure naive Bayes classification approach. The [13] presents another hybrid method of naïve Bayes with self organizing map (SOM). Proposed Bayes classifier used at the front end, while SOM performs the indexing steps to retrieve the best match cases.

A hybrid algorithm is proposed in [14], based on variable precision rough set to combine the strength of both k-NN and Rocchio techniques to improve the text classification accuracy and overcome the weaknesses of Rocchio algorithm.

So In the context of combining multiple classifiers for text categorization, a number of researchers have shown that combining different classifiers can improve classification accuracy. It is observed from the Comparison between the best individual classifier and the combined method, that the performance of the combined method is superior.

TABLE I: Comparative Results among Different Classifiers Obtained on Reuters 21578 and 20 Newsgroup Datasets [19]

Results reported by	Dataset	Representation Scheme	Classifier Used	Macro F1	Micro F1
[Ko et al., 2004] [5]	20 Newsgroup	Vector representation with different weights	Naïve Bayes	83.00	83.00
			Rocchio	78.60	79.10
			K-NN	81.20	81.04
			SVM	86.00	86.10
[Tan et al., 2005] [6]	20 Newsgroup	Vector representation	Naïve Bayes	0.835	0.835
			Centroid	0.838	0.842
			K-NN	0.846	0.848
			SVM	0.887	0.889
[Liang et al., 2005] [7]	Reuters 21578	Vector representation	K-NN	-	0.797
[Mubaid and Umair, 2006] [8]	20Newsgroup	Vector representation	L Square	83.05	86.45
			SVM	78.19	84.62
	Reuters 21578	Vector representation	L Square	94.57	-
			SVM	95.53	-
[Hao et al., 2006] [9]	Reuters 21578	Hierarchical graph structure	SVM(Polynomial)	-	86.20
			SVM(rbf)	-	86.50
			K-NN	-	0.7888
			Decision Tree	-	0.879
[Lan et al., 2009] [10]	Reuters 21578	VSM with term weighting schemes	SVM	0.900	0.921
			K-NN	0.825	0.840
	20Newsgroup	VSM with term weighting schemes	SVM	0.808	0.808
			K-NN	0.691	0.691

VI. Conclusion

Currently text categorization research is investigating the scalability properties of text classification systems, i.e. understanding whether the systems that have proven the best in terms of effectiveness alone stand up to the challenge of dealing with very large numbers of categories.

Several algorithms or combination of algorithms as hybrid approaches were proposed for the automatic classification of documents. Among these algorithms SVM, NB, kNN and their hybrid system with the combination of different other algorithms and feature selection techniques are shown most appropriate in the existing literature.

Future work is required for the performance improvement and accuracy of the text classification process.

After performing a review on different types of approaches and comparing existing methods based on various parameters it can be concluded that SVM classifier has been recognized as one of the most effective text classification method in the comparisons of supervised machine learning algorithms.

References

[1] Tam, Santoso A and Setiono R., "A comparative study of centroid-based, neighborhood-based and statistical approaches for effective document categorization", *ICPR '02 Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02)* ,vol.4 , no. 4 , 2002, pp.235–238.

[2] William Cohen and Yoram, "Context-sensitive learning method for text categorization", *Proc. of SIGIR 96, 19th International Conference on Research and Development in Informational Retrieval*, vol. 17, Issue 2, April 1999 ,pp-307-315.

[3] Russell Greiner and Jonathan Schaffer, "Exploratorium – Decision Trees", Canada. 2001. *URL: http://www.cs.ualberta.ca/~aixplore/ learning/ Decision Trees.*

[4] Irina Rish, "An Empirical Study of the Naïve Bayes Classifier", Proc. of the IJCAI-01 *Workshop on Empirical Methods in Artificial Intelligence*, Oct 2001. citeulike-article-id:352583.

[5] KO, Y. J., Park, J., and Seo, J., "Improving text categorization using the importance of sentences", *International Journal Information Processing and Management*, vol. 40, no. 1, January 2004, pp. 65-79.

[6] Songbo, T., Cheng, X., Ghanem, M. M., Wnag, B. and Xu, H., "A novel refinement approach for text categorization", *Proc. of 14th ACM International Conference on Information and Knowledge Management*, 2005, pp.469-476.

[7] Liang, C. Y., Guo, L., Xia, Z. H., Nie, F. G., Li, X. X., Su, L., and Yang, Z. Y. , "Dictionary-based text categorization of chemical web pages", *International Journal Information Processing and Management*, vol. 42, no. 4, July 2006, pp.1072 – 1029.

[8] Mubaid H. A, and Umair 2006, "A New Text Categorization Technique using Distributional Clustering and Learning Logic", IEEE Trans. on Knowledge and Data Engineering, vol.18, no..9, September 2006, pp. 1156 – 1165.

[9] Hao, P. Y., Chaing, J. H., and Tu, Y. K.., "Hierarchically SVM classification based on support vector clustering method and its application to document categorization", *International Journal Expert Systems with Applications*, vol. 33, no. 3, October 2007, pp. 1-5.

[10] Thorstan J., "Text Categorization with Support Vector Machins Learning with many relevant features" , vol. 1398, 1998, pp.137-142.

[11] Tan, C. L., Su. J., and Lu, Y., "Supervised and Traditional Term Weighting Methods for Automatic Text Categorization", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no 4, April 2009, pp. 721 – 735.

[12] Dino Isa, Lam Hong lee, V. P Kallimani, and R. Raj Kumar, "Text Documents Preprocessing with the Bahes Formula for Classification using the Support vector machine", *IEEE Trans. of Knowledge and Data Engineering*, vol.20, no. 9, September 2008, pp.1264-1272.

[13] Dino Isa, and V. P Kallimani Lam Hong lee, "Using Self Organizing Map for Clustering of Text Documents", *Elsevier , Expert System with* Applications, vol. 36, no. 5, July 2009, pp. 9584-9591.

[14] Duoqian Miao , Qiguo Duan, Hongyun Zhang, and Na Jiao, "Rough set based hybrid algorithm for text classification",Journal of Expert Systems with Applications, vol. 36, no. 5, July 2009, pp. 9168-9174..

[15] Taeho Jo," NTC (Neural Text Categorizer): Neural Network for Text Categorization", *International Journal of Information Studies*, vol. 2, no.2, April 2010.

[16] Patrick Marques Ciarelli, Elias Oliveira, Claudine Badue and Alberto Ferreira De Souza, "Multi-Label Text Categorization Using a Probabilistic Neural Network" *International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM),* vol.1 ,2009, pp.133-144 ,.

[17] B. Mahalakshmi, K.Duraiswamy, and Tintu Paul, "Improving the Efficiency of Text Categorization by Self-Organizing Map", *Proceedings of the Int. Conf. on Information Science and Applications ICISA 2010*, 6 February 2010, Chennai, India.

[18] S.Ramasundaram, "Text Categorization by Backpropagation Network", , International Journal of Computer Applications, Vol., no.6, October 2010, pp1-5.

[19] S. Manjunath, B.S. Harish, "Representation and Classification of Text Documents : A Brief Review" *IJCA Special Issue on "Recent Trends in Image Processing and Pattern Recognition*, RTIPPR,2010, pp. 110-119.

Pratiksha P. Pawar was born on 19th December 1983. She received B. E. and M. Tech. degrees in computer engineering from College of Engineering Pune, University of Pune in 2005 and 2009 respectively.

Currently She is working as Assistant Professor in Computer Engineering department of Government College of engineering and Research, Awasari, Pune. She is life member of ISTE.

S. H. Gawande was born on 4th July 1979 in small village Deori Tq. Akot, Dist. Akola in Maharashtra State. He completed B.E. degree in mechanical engineering from Amravati University, Amravati in April 2001 and M.E. degree in mechanical engineering with design engineering as specialization in December 2002 from University of Pune. Now he is working as Assistant Professor in mechanical engineering at M.E.S. College of Engineering Pune, India from 2004. His research interests include internal combustion engines, design engineering, and Tribology. He is permanent member of Indian societies like ISTE from 2005, SAE from 2008 and IACSIT Singapore from 2009.