International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-10, August 2019

Text Classification Using Ensemble Of Non-Linear Support Vector Machines

Sheelesh Kumar Sharma, Navel Kishor Sharma

Abstract: ***With the advent of digital era, billions of the documents generate every day that need to be managed, processed and classified. Enormous size of text data is available on world wide web and other sources. As a first step of managing this mammoth data is the classification of available documents in right categories. Supervised machine learning approaches try to solve the problem of document classification but working on large data sets of heterogeneous classes is a big challenge. Automatic tagging and classification of the text document is a useful task due to its many potential applications such as classifying emails into spam or non-spam categories, news articles into political, entertainment, stock market, sports news, etc. The paper proposes a novel approach for classifying the text into known classes using an ensemble of refined Support Vector Machines. The advantage of proposed technique is that it can considerably reduce the size of the training data by adopting dimensionality reduction as pre-training step. The proposed technique has been used on three bench-marked data sets namely CMU Dataset, 20 Newsgroups Dataset, and Classic Dataset. Experimental results show that proposed approach is more accurate and efficient as compared to other state-of-the-art methods.***

Keywords: Text classification, support vector machine, non-linear ensemble, machine learning, natural language processing.

I. INTRODUCTION

With the advent of digital era, billions of the documents generate every day that need to be managed, processed and classified. Enormous size of text data is available on world wide web and other sources. As a first step of managing this mammoth data is the classification of available documents in right categories. Supervised machine learning approaches try to solve the problem of document classification but working on large data sets of heterogeneous classes is a big challenge. Automatic tagging and classification of the text document is a useful task due to its many potential applications such as classifying emails into spam or non-spam categories, news articles into political, entertainment, stock market, sports news, etc. The paper proposes a novel approach for classifying the text into known classes using an ensemble of refined Support Vector Machines. The advantage of proposed technique is that it can considerably reduce the size of the training data by adopting dimensionality reduction as pre-training step. The proposed technique has been used on three bench-marked data sets namely CMU Dataset, 20 Newsgroups Dataset, and Classic Dataset. Experimental results show that proposed approach is more accurate and efficient as compared to other state-of-the-art methods.

II. LITERATURE SURVEY

The area of text mining has been popular among researchers for quite a long time. In a classic survey work of Berry [1], clustering, classification and retrieval of text data have been discussed along with various other concepts of text mining. Hotho et al [2] also present a survey on text mining along with various pre-processing steps and algorithms. Text classification has many interesting applications such as content management, fraud detection in banking, sentiment analysis, customer reviews and feedback analysis, search engine optimization, biomedical analysis etc[3]-[7].

Text classification is a supervised learning task. Many approaches have been deployed for performing it. Traditionally, the approaches that can be found in literature for text classification include naive Bayes classifier, k-nearest neighbors, artificial neural network, evolutionary approaches, support vector machines, decision trees etc [8]-[11]. The training of the classifier can be either feature based or end-to-end learning without the need of the step of feature extraction. Provided with the huge volume of data, dimensionality reduction step can substantially reduce its size. There are many approaches for dimensionality reduction. Two popular approaches are linear discriminant analysis (LDA) and principal component analysis (PCA). It is always computational efficient to work with reduced data as compared to the entire data in raw form.

Deep learning based methods have been very effective especially in visual and textual pattern recognition tasks[12]-[13]. These approaches can be either feature based or can use end-to-end learning without the need of feature extraction. The end-to-end learning variant of deep learning is highly popular. One limitation with deep learning based methods is that they require lots of data and computation resources. There are many deep learning models. The most popular model is convolutional neural network (CNN). To overcome the shortcomings of CNN, some advanced models like recurrent neural network (RNN), long short term memory (LSTM) network- a variant of RNN etc[12]-[13] exist. Still, there is one common limitation with deep learning methods that they require huge training data and the computation resources. Support vector machines have been a popular choice for training binary classifiers.

Manuscript published on 30 August 2019.
***Correspondence Author(s)**
Dr Sheelesh Kumar Sharma, Professor (Comp. Sc.), IMS Ghaziabad, India.
Mr Navel Kishor Sharma, Associate Dean, Academic City College Ghana.

III. WHAT IS SUPPORT VECTOR MACHINES AND ENSEMBLE OF SUPPORT VECTOR MACHINES

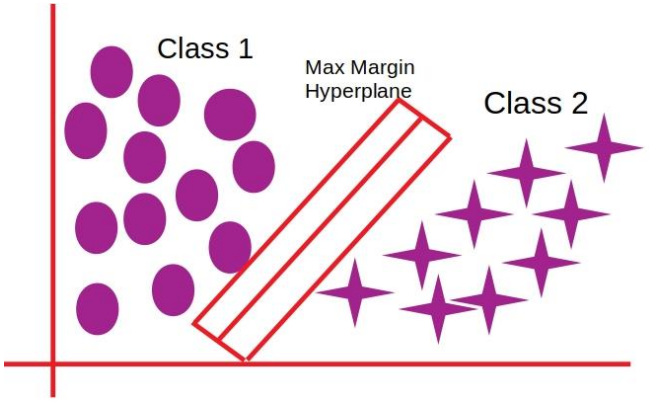

Figure 1: A Linear Support Vector Machine

Figure 1 shows a linear support vector machine having two classes namely class 1 and class 2. For classification a hyperplane is found that separates these two classes maximizing the distance from the support vector of each classes. The extreme feature points placed on boundary each class are called support points and form a support vector. The data which is not linearly separable can be classified with SVM using kernel method.

There exist various variants of SVMs[14]-[19]. These are the state-of-the-art SVM models that have been used for text classification. Wang et al[15] propose a fuzzy SVM based approach for text classification. They take the advantage of fuzzy logic and overfitting resistance due to SVM. SVM are good for solving multi-class classification and regression problems as suggested in [16]-[18]. The power of modern computing such as parallel processing can be harnessed with SVM classier. In a recent work of Chatterjee et al.[17], multithreading and CUDA have been used for reducing time and achieving computational efficiency.

Goudjil et al.[18] use a novel active learning approach with SVM for text classification. SVM based classifiers have proved to be a viable option for larges scale text classification problems. In the work of Do et al[19], a latent SVM based text classification approach is proposed that works fine with large data sets.

IV. ADVANTAGES OF SUPPORT VECTOR MACHINE OVER OTHER CLASSIFIERS

The motivation behind using support vector machines in the present work can be described in the following points:

i. Overfitting is a concern for classifier training. Support vector machine is almost away from overfitting.

ii. They can be used for various forms of data such as semi-structured and unstructured data including text, images, numeric values etc.

iii. SVM models are easily scalable

iv. SVM models are easy to train and give better results than various other complex models such as ANN

v. Support vector machines are good at linear classification as well as they can be used for non-linear classification using kernels.

V. ENSEMBLES OF SUPPORT VECTOR MACHINES

Ensembles of the support vector machines are good to achieve a better classification accuracy in multi-class classification problems. In ensemble classification, the results received from various weak base learners are aggregated. Different techniques are available to create ensemble of classifiers that can be termed as bagging or boosting.

The popular ensemble constructing techniques that are used for text classification task include AdaBoost, Arc-X4, modified Adaboost etc. Bagging is an aggregation algorithm used to improve the stability and accuracy of classification techniques. Bagging happens to be the shortened form for Bootstrap Aggregating. It is closely associated with decision tree classifiers but now it can be used with any type of classification or regression algorithm.

Wang et al [14], present an exhaustive empirical study of ensemble techniques for support vector machines. They evaluate these techniques on twenty data sets taken from UCI repository. SVM ensembles may correspond to the cross-validation optimization of single SVM. One notable advantage of ensemble based classification approach is the stable classification performance than other models.

VI. PROPOSED METHOD FOR TEXT CLASSIFICATION

The basic idea behind support vector machine is to fit a hyperplane or kernel that can discriminate one feature type from another in such a way that the distance between different feature points is maximized. The proposed methodology harnesses the power of multiple support vector machines as binary classifiers to construct a more viable multi-class classifier. Supervised training is performed on the corpus of training data and then testing is performed on the test set of data.

To accommodate the large size of training set, most of the common words are filtered out (stopping words). This step reduced the size of data (word count) on an average up to 40%. Instead of performing training on the raw data, it is a good to extract the features from data first, and then perform training using feature vector. To reduce its size further, LDA (linear discriminant analysis) is used.

Figure 2 depicts the schematic diagram of the proposed method of text document classification. The proposed technique can be broken down into five basic steps.

Retrieval Number: J95200881019/19©BEIESP
DOI: 10.35940/ijitee.J9520.0881019
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

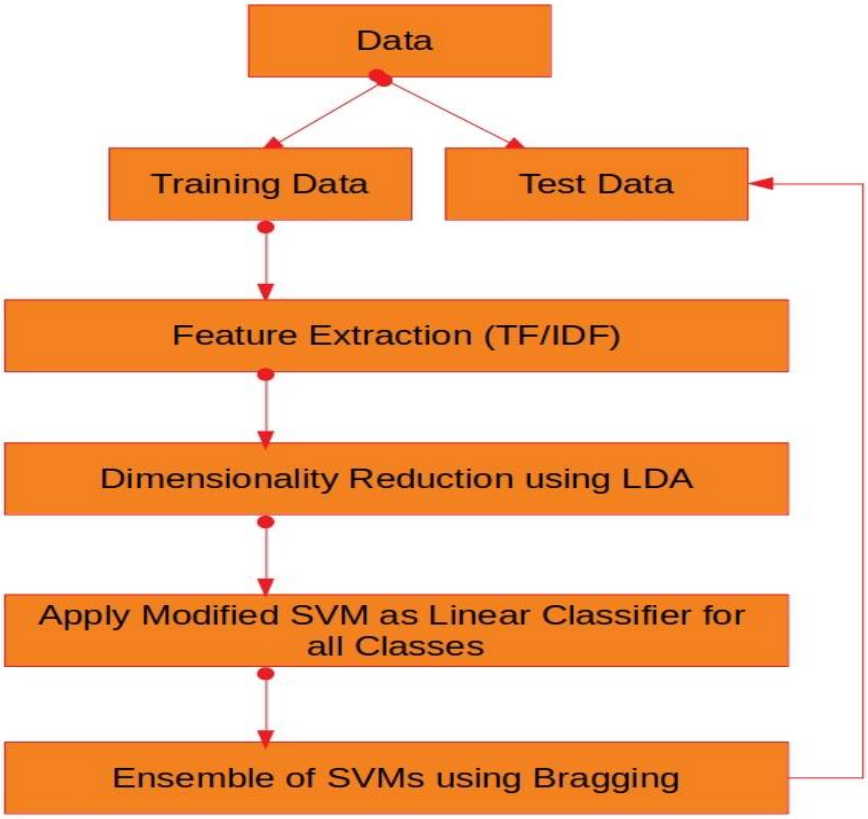

Figure 2: Proposed Method for Text Document Classification

Step 1: Train and validation split

Available training corps is divided into two parts called as training data and validation data. The training set is used to train the classifier and validation set is used to estimate the accuracy of the classifier. In order to make the process more effective, the samples need to be selected randomly. We divide the data in the ratio of 80:20 i.e 80% data for training and 20% for validation.

Step 2: Feature Extraction using TF/IDF

Term Frequency (TF) and Inverse Term Frequency are the popular feature representation metrics that normalize the importance of a word and its association with a particular class of the text. TF-IDF value is directly proportional to the occurrence of a word in a text document. Calculation of TF-IDF value is simple.

Here, we use a modified square rooted TF-IDF metric. This metric improves the purity and entropy of the classification results and avoids skewedness towards mean error rate.

Where $sqwv_i$ represents frequency of i^{th} word in the document. Analogous to word frequency, the term frequency (tf) can be calculated as follows:

$$tf(t,d) = 0.5 + 0.5xf_{t',d}\,(max f_{t'},_d : t' \in d) \quad (1)$$

Here, we use square root value of the term frequency. The modified term frequency is can be used to represent a document as follows:

$$\boldsymbol{DOCUMENT_i} = (sqtf_1, sqtf_2, \ldots\ldots sqtf_n) \quad (2)$$

Inverse Document Frequency (IDF) measures how uniquely a document a document lies across rest of the other documents. It is calculated by the following method:

IDF= log (total documents in corps/ number of documents having a given term)

$$idf(t,d) = log\,\frac{N}{[mode\,(d \in D : t \in D)]} \quad (3)$$

Step 3: Dimesionality Reduction using LDA

When working with very large data, the dimensionality reduction step can substantially reduce the data to be processed without compromising the end results. Various techniques exist for performing this task. LDA and PCA are commonly used for dimensionality reduction. In our proposed technique, we use LDA. LDA projects the training dataset into the new feature space of reduced dimensions.

In LDA approach, we try to maximize the function that represents the difference between the means (averages). The means are normalized by a metric of the within-class variability. LDA can also be used a linear classier as well as a tool for dimensionality reduction. LDA calculates

centroid of each class in the feature set. Suppose if there are 20 different feature sets, then LDA will calculate the centroid of every class. Further, it will re-project the feature points to a new dimensions.

For doing this a new axis is calculated satisfying the following two objectives:

i. The centroids of the classes should be at maximum distance.

ii. Minimize the variation within each category

Step 4: Non-Linear SVM Kernel

Here, we are considering the case of non-linearly separable data points. Linear SVM cannot classify the data which is non-linear in nature. There exist alternatives to linear SVM that can help in classification of linearly non-seperable data. One way is projecting the data points to higher dimensions i.e x to x^2, a polar coordinate projection may be another possibility. In practical life, most of the times the data that we encounter is randomly distributed.

To classify linearly non-separable data, SVM uses a kernel trick which helps to use a linear classifier on non linear data. A variety of kernels can be used for this purpose such as Sigmoid, Polynomial, RBF (Radial Basis Function) kernels. Here, we use RBF Kernel for high dimension projection. The formula for calculating RBF kernel of two feature points x and y can is represented in the form of radial basis functions $\phi(x)$ and $\phi(y)$ as follows:

$$K(x, y) = \phi(x)^T \phi(y)^T$$

There is no need of calculating $\phi(x)$ and $\phi(y)$. It can be reduced to the simple expression:

$$K(x, y) = exp(-\gamma(x - y)^2), \gamma > 0$$

Where γ is a constant.

Step 5: Ensemble of Modified SVMs using Bagging

Ensemble means combining various classifiers into one for performing a given task. Here, objective is to combine binary SVMs to combine into a single multi-class SVM that can classify the text document into one of the known categories after training. Bagging is a technique for ensemble creation and it stands for "Bootstrap Averaging".

Let there be a training set of size *n*. With the help of bagging, we generate *m* new training sets of size n′ each. A uniform replacement sampling is done with around (1 - 1/*e*) fraction of uniqueness. It results into about 63% unique samples and rest being duplicate. It is called bootstrap sample. *m* bootstrap samples fit *m* models (SVMs). Further, they are combined by voting in case of classification. For regression, they can be simply averaged. One advantage of using bagging is that it overcomes the problem of over-fitting.

VII. EXPERIMENTAL RESULTS

The proposed techniques has been tested on three bench-marked datasets. These data sets are are CMU Knowledge-base Dataset [20], 20 Newsgroups Dataset [21], and Classic Dataset [22]. These datasets are from distinctive document categories. 20 Newsgroups Dataset has the collection of 18828 documents of 20 categories.

Classic Dataset is the collection of research papers from 4 different disciplines. Lastly, CMU Web Knowledge-base dataset is the collection of 8282 web pages of 7 different categories.

Table 1: Description of the Datasets used for Evaluation of Proposed Methodology

S.No	*Name of Dataset*	*# Documents*	*#Classes*	*#Terms*	*Avg Class Size*	*Type of Documents*
1	CMU Knowledge-base Dataset [20]	8282	07	20682	1050	Collection of web pages
2	20 Newsgroups Dataset [21]	18828	20	28553	1217	Newsgroup post data
3	Classic Dataset [22]	7095	4	12009	1774	Academic papers falling under 4 categories. **CACM**: 3204 documents **CISI**: 1460 documents **CRAN**: 1398 documents **MED**: 1033 documents

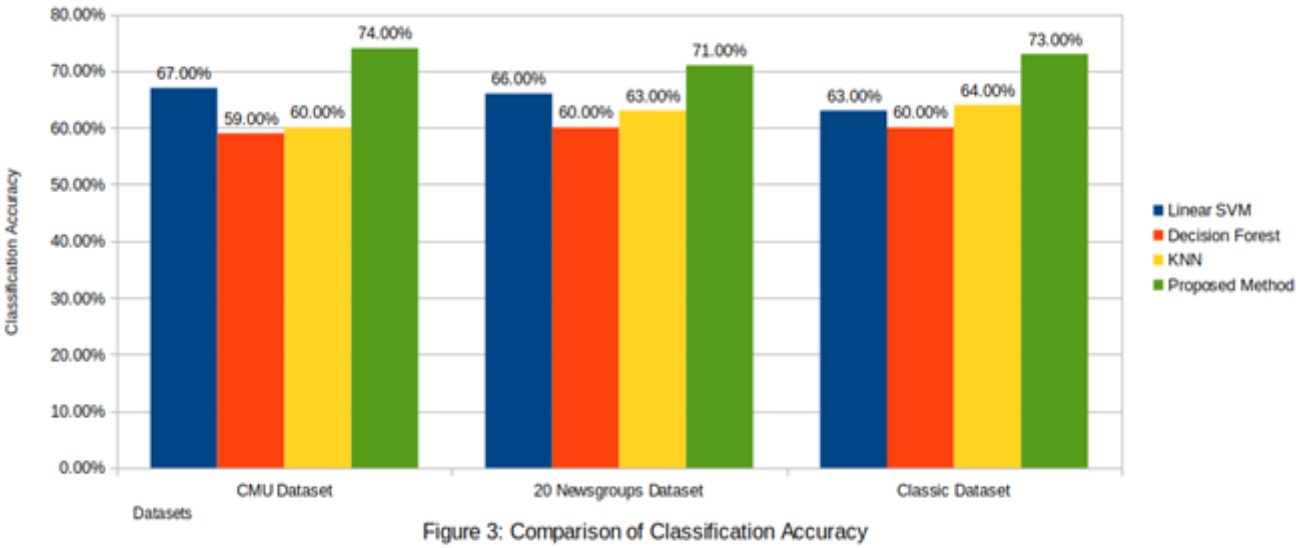

Figure 3: Comparison of Classification Accuracy

VIII. COMPARISON WITH OTHER METHODS

The proposed method has been compared with three other classification techniques namely linear SVM, Decision Forest (random forest), and KNN on three different data sets[20]-[22]. The comparison of classification accuracy has been shown in the figure 3. It is evident that the proposed technique outperforms all the three methods. Proposed method has the average classification accuracy of 72.66%.

IX. CONCLUSION

The problem of text classification has been discussed in the paper. Text classification is a challenging problem. Many potential applications make it important. A supervised learning approach with ensemble of non-linear support vector machines has been proposed in the paper. Elimination of trivial features is performed with LDA. Experimental results show that proposed technique is better than other classification methods including decision forest (random forest), linear SVM and KNN on three bench-marked datasets.

REFERENCES

1. Berry, M. W. (2004). Survey of text mining. Computing Reviews, 45(9), 548.
2. Hotho, A., Nürnberger, A., & Paaß, G. (2005, May). A brief survey of text mining. In Ldv Forum (Vol. 20, No. 1, pp. 19-62).
3. Kumar, B. S., & Ravi, V. (2016). A survey of the applications of text mining in financial domain. Knowledge-Based Systems, 114, 128-147.
4. Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., ... & Shen, B. (2013). Biomedical text mining and its applications in cancer research. Journal of biomedical informatics, 46(2), 200-211.
5. AlSumait, L., Barbará, D., & Domeniconi, C. (2008, December). On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In 2008 eighth IEEE international conference on data mining (pp. 3-12). IEEE.
6. He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. International Journal of Information Management, 33(3), 464-472.
7. Zanasi, A., & Zanasi, A. (2007). Text mining and its applications to intelligence, CRM and knowledge management. Wit Press.
8. Kang, M., Ahn, J., & Lee, K. (2018). Opinion mining using ensemble text hidden Markov models for text classification. Expert Systems with Applications, 94, 218-227.
9. Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In Advances in neural information processing systems (pp. 649-657).
10. Conneau, A., Schwenk, H., Barrault, L., & Lecun, Y. (2016). Very deep convolutional networks for text classification. arXiv preprint arXiv:1606.01781.
11. Skrlj, B., Kralj, J., Lavrač, N., & Pollak, S. (2019). Towards Robust Text Classification with Semantics-Aware Recurrent Neural Architecture. Machine Learning and Knowledge Extraction, 1(2), 575-589.
12. Abdi A., Shamsuddin S.M, Hasan S., & Piran J. (2019): Deep learning based sentiment classification of evaluative text based on multi-feature fusion, Information Processing and Management, 56, 1245-1259.
13. He J., Wang L., Liu L., Feng J., & Wu H. (2019) Long Document Classification From Local Word Glimpses via Recurrent Attention Learning, 7, 40707 – 40718 .
14. Wang, S. J., Mathew, A., Chen, Y., Xi, L. F., Ma, L., & Lee, J. (2009). Empirical analysis of support vector machine ensemble classifiers. Expert Systems with applications, 36(3), 6466-6476.
15. Wang, T. Y., & Chiang, H. M. (2007). Fuzzy support vector machine for multi-class text categorization. Information Processing & Management, 43(4), 914-929.
16. Tang, F. M., Wang, Z. D., & Chen, M. Y. (2005). On multiclass classification methods for support vector machines. Control and Decision, 20(7), 746.
17. Chatterjee, S., Jose, P. G., & Datta, D. (2019). Text Classification Using SVM Enhanced by Multithreading and CUDA. International Journal of Modern Education and Computer Science, 11(1), 11.
18. Goudjil, M., Koudil, M., Bedda, M., & Ghoggali, N. (2018). A novel active learning method using SVM for text classification. International Journal of Automation and Computing, 1-9.
19. Do, T. N., & Poulet, F. (2019). Latent lSVM classification of very high dimensional and large scale multi-class datasets. Concurrency and Computation: Practice and Experience, 31(2), e4224.
20. CMU Web Knowledgebase Dataset http://www.cs.cmu.edu/afs /cs.cmu.edu/project/theo-20/www/data/ (last visited on June 15, 2019)
21. 20 News Net Dataset: http://qwone.com/~jason/20Newsgroups/ (last visited on June 15, 2019)
22. Classic dataset http://www.dataminingresearch.com/ download/ dataset/classicdocs.rar (last visited on June 15, 2019)

AUTHORS' PROFILE

Dr Sheelesh Kumar Sharma is a Professor in MCA Department at IMS Ghaziabad. He obtained **MCA** from M. B. M. Engg. College Jodhpur, **M. Tech.** in Computer Science from Institution of Electronics and Telecommunication Engineers New Delhi and **M. Phil.** in Computer Science from Madurai Kamraj University Madurai and He has awarded **Ph.D.** from **University of Rajasthan**, **Jaipur** and his research area is data mining and data warehousing. He has more than 18 years of experience in academics and 6 years of Industrial experience. He is a life time member of the Institution of Electronics and Telecommunication Engineers (IETE) Delhi.

Navel Kishore Sharma is an Associate Dean at Academic City College Accra in Ghana. He did his M. Tech. from Rajasthan University, Jaipur in 2006 and pursuing PhD. He has more than **25 years** of professional experience in the fields of Software Development and Teaching in Engineering Colleges.

Retrieval Number: J95200881019/19©BEIESP
DOI: 10.35940/ijitee.J9520.0881019
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)