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.
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.

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

![Figure 1: A Linear Support Vector Machine](image)

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 class. 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 classifier. In a recent work of Chatterjee et al.[17], multithreading and CUDA have been used for reducing time and achieving computational efficiency.

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

$$tf(t, d) = 0.5 + 0.5xf_{i,d}(\max_{f_{i,d}':t' \in d}) \tag{1}$$

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

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

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:

$$IDF = \log \frac{\text{total documents in corps}}{\text{number of documents having a given term}} \tag{3}$$

**Step 3: Dimensionality 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 classifier 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-separable 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:

\[
K(x, y) = \phi(x)^T \phi(y)^7
\]

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 \( \gamma \) 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 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>
<thead>
<tr>
<th>S.No</th>
<th>Name of Dataset</th>
<th># Documents</th>
<th>#Classes</th>
<th>#Terms</th>
<th>Avg Class Size</th>
<th>Type of Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CMU Knowledge-base Dataset [20]</td>
<td>8282</td>
<td>07</td>
<td>20682</td>
<td>1050</td>
<td>Collection of web pages</td>
</tr>
<tr>
<td>2</td>
<td>20 Newsgroups Dataset [21]</td>
<td>18828</td>
<td>20</td>
<td>28553</td>
<td>1217</td>
<td>Newsgroup post data</td>
</tr>
<tr>
<td>3</td>
<td>Classic Dataset [22]</td>
<td>7095</td>
<td>4</td>
<td>12009</td>
<td>1774</td>
<td>Academic papers falling under 4 categories.</td>
</tr>
</tbody>
</table>

CACM: 3204 documents
CISI: 1460 documents
CRAN: 1398 documents
MED: 1033 documents

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

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.