International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 2, February - 2014

A Review on Feature Selection and Document Classification using Support Vector Machine

Ms. Yoginee R. Surkar
M.Tech III sem,
Dept. of Computer Science and Engineering
Bapurao Deshmukh College of Engineering
Sevagram, Maharashtra, India

Prof. S. W. Mohod
Dept. of Computer Science and Engineering
Bapurao Deshmukh College of Engineering
Sevagram, Maharashtra, India

Abstract **- Recently the development of high performance, automatic text classification as well as feature selection has very challenging and tedious process because of many problems have been still unresolved. This paper provide the work done so far for the development of most efficient technique on the feature selection and document classification using support vector machine (SVM). Feature selection improves the efficiency and accuracy of text classification algorithm by removing redundant and irrelevant terms. Instead of the work done yet there is need to achieve huge level accuracy by reducing the computational time.**

Keywords- feature selection, text classification, SVM.

1. INTRODUCTION

Support Vector Machine (SVM), is one of best machine learning algorithms, which was proposed in 1990's and used mostly for pattern recognition. This has also been applied to many pattern classification problems such as image recognition, speech recognition, text categorization, face detection and faulty card detection, etc. Pattern recognition aims to classify data based on either a priori knowledge or statistical information extracted from raw data, which is a powerful tool in data separation in many disciplines. SVM is a supervised type of machine learning algorithm in which, given a set of training examples, each marked as belonging to one of the many categories, an SVM training algorithm builds a model that predicts the category of the new example. SVM has the greater ability to generalize the problem, which is the goal in statistical learning.

1.1 SVM Model

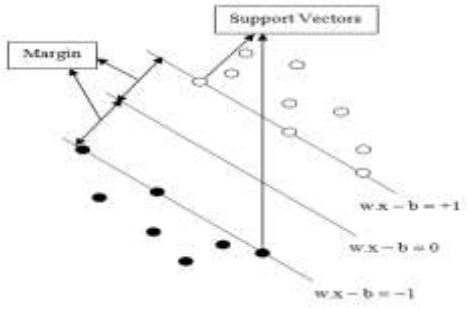

Figure 1: SVM model

The figure 1 is the simple model for representing support vector machine technique. The model consists of two different patterns and the goal of SVM is to separate these two patterns. The model consists of three different lines. The line w.x-b=0 is known as margin of separation or marginal line.

The lines w.x - b = 1 and w.x – b = -1 are the lines on the either side of the line of margin. These three lines together construct the hyper plane that separates the given patterns and the pattern that lies on the edges of the hyper plane is called support vectors. The perpendicular distance between the line of margin and the edges of hyper plane is known as margin. One of the objectives of SVM for accurate classification is to maximize this margin for better classification. The larger the value of margin or the perpendicular distance, the better is the classification process and hence minimizing the occurrence of error [11].

1.2 Feature Selection

Today people are facing an overflow of data in digital format, for example, vast volumes of online text are available via the World Wide Web (WWW), and news feeds, electronic mails, teaching materials and so on. At present, documents classifications are done manually in many companies. Though significant resources have been spent on it, the manual classification is becoming an impossible mission as the number of documents rapidly increases over time. Developing the automatic classifier which can be used to accurately and efficiently classify the existing and incoming documents is of great importance, which is very useful for many text-based applications such as web categorization, information retrieval, and email sort [14].

Document classification can be defined as the task of automatically categorizing collections of electronic documents into their annotated classes based on their contents. In recent years, this has become important due to the advent of large amounts of data in digital form. For several decades now, document classification in the form of text classification systems have been widely implemented in numerous applications such as spam filtering, e-mails

categorizing, directory maintenance, and ontology mapping. An increasing number of statistical and computational approaches have been developed for document classification, including k-nearest- neighbor (k-NN) classification, naive Bayes classification, support vector machines (SVMs), maximum entropy, decision tree induction, rule induction, and artificial neural networks [3].

Developing the high performance automatic text classification method is very challenging and many problems are still unresolved. This mainly focuses on one of these difficult problems, the implementation of the feature selection. Vector Space Model (VSM) is widely used in the text classifying systems to transfer the unstructured text data into structured ones. It treats a document as a bag of words. In fact, not all features are useful for constructing the document classifier. Some of the features may be redundant or irrelevant. Furthermore, some may even misguide the classifying result, especially when there are more irrelevant features than relevant ones. In such case, selecting a subset of original features often leads to a better classifying performance. A good feature selection method is not only important for the efficiency and but also the accuracy of the text classification.

Feature selection can improve the efficiency and accuracy of text classification algorithms by removing redundant and irrelevant terms. Some existing feature selection algorithms shown that the DF (Document Frequency) algorithm is the simplest method and can obtain the competitive results compared with other much more complicated algorithms[14].

2. RELATED WORK

Z. Zheng and R. Srihari [1] in (2003) proposed a novel local feature selection approach for text categorization. It constructs a feature set for each category by first selecting a set of terms highly indicative of membership as well as another set of terms highly indicative of non-membership, then unifying the two sets. The size ratio of the two sets was empirically chosen to obtain optimal performance. This is in contrast with the standard local feature selection approaches that either only selects the terms most indicative of membership; or implicitly but not optimally combine the terms most indicative of membership with non-membership. The comparison between the proposed approach and standard approaches was conducted on four feature selection metrics: chisquare, correlation coefficient, odds ratio, and GSS coefficient. The results show that the proposed approach improves text categorization performance.

Anirban Dasgupta, Petros Drineas, Boulos Harb, Vanja Josifovski and Michael W. Mahoney [2] in (2007) proposed an unsupervised feature selection strategy for that they give worst-case theoretical guarantees on the generalization power of the resultant classification function ˜ f with respect to the classification function f obtained when keeping all the features. This is the first feature selection method with such guarantees. In addition, the analysis leads to insights as to when and why this feature selection strategy will perform well in practice. They used the TechTC-100, 20-Newsgroups, and Reuters-RCV2 data sets to evaluate empirically the performance of this and two simpler but related feature selection strategies against two commonly-used strategies. Their empirical evaluation shows that the strategy with provable performance guarantees performs well in comparison with other commonly-used feature selection strategies. In addition, it performs better on certain datasets under very aggressive feature selection.

Dino Isa, Lam Hong Lee, V.P. Kallimani, and R. RajKumar [3] in (2008) implemented an enhanced hybrid classification method through the utilization of the naive Bayes approach and the Support Vector Machine (SVM). The Bayes formula was used to vectorized a document according to a probability distribution reflecting the probable categories that the document may belong to. The Bayes formula gives a range of probabilities to which the document can be assigned according to a predetermined set of topics (categories) such as those found in the "20 Newsgroups" data set for instance. Using this probability distribution as the vectors to represent the document, the SVM can be used to classify the documents on a multidimensional level. The effects of an inadvertent dimensionality reduction caused by classifying using only the highest probability using the naive Bayes classifier can be overcome using the SVM by employing all the probability values associated with every category for each document. This method can be used for any data set and shows a significant reduction in training time as compared to the Lsquare method and significant improvement in the classification accuracy when compared to pure naive Bayes systems and also the TF-IDF/SVM hybrids.

Janez Brank, Dunja Mladenić, Marko Grobelnik and Natasa Milic-Frayling [4] in (2008) used feature selection methods that are applied in the context of document classification. Those are important for processing large data sets that may contain millions of documents and are typically represented by a large number, possibly tens of thousands of features. Processing large data sets thus raises the issue of computational resources and they find the right trade-off between the size of the feature set and the number of training data that can take into account. Depending on the selected classification technique, different feature selection methods require different optimization approaches, raising the issue of compatibility between the two. They demonstrate effective classifier training and feature selection method that is suitable for large data collections. They explore feature selection based on the weights obtained from linear classifiers themselves, trained on a subset of training documents. They show how these feature selection methods combine with various learning algorithms. Their experiments include a comparative analysis of three learning algorithms: Naive Bayes, Perceptron, and Support Vector Machines (SVM) in

combination with three feature weighting methods: Odds ratio, Information Gain, and weights from the linear SVM and Perceptron. They show that by regulating the size of the feature space using an effective feature scoring, like linear SVM, need only a half or even a quarter of the computer memory to train a classifier of almost the same quality as the one obtained from the complete data set. Feature selection using weights from the linear SVMs yields a better classification performance than other feature weighting methods when combined with the three learning algorithms.

Debnath Bhattacharyya, Poulami Das, Debashis Ganguly, Kheyali Mitra, Purnendu Das, Samir Kumar Bandyopadhyay and Tai-hoon Kim [5] in (2009) proposed a technique for unstructured document categorization. That means, the collected unstructured documents will be categorized based on some given constraints. It deals with different techniques like text and data mining, genetic algorithm, lexical chaining, binarization method to reach the fulfillment of desired unstructured document categorization.

Atika Mustafa, Ali Akbar, and Ahmer Sultan [6] in (2009) implemented Information Extraction and Categorization in the text mining application. Textual data in electronic documents today around the world have brought forward all the information one could need and as data banks build up worldwide, and access gets easier through technology, it has become easier to overlook vital facts and figures that could bring about groundbreaking discoveries. To extract terms from the document they used modified version of Porter's Algorithm for inflectional stemming. For calculating term frequencies for categorization, they used a domain dictionary for 'Computer Science' domain.

Othman, M. S.,Yusuf, L. M and Salim J [7] in (2010) discussed the web document features that classify the web information resources. Six web document features have been identified which are text, meta tag and title (A), title and text (B), title (C), meta tag and title (D), meta tag (E) and text (F). The Support Vector Machine (SVM) method is used to classify the web document while four types of kernels namely: Radial Basis Function (RBF), linear, polynomial and sigmoid kernels was applied to test the accuracy of the classification. The studies show that the text, meta tag and title (A) features is the best features for classification of web document that employs the four kernels followed by the features on title and text (B) as well as the features on meta tag and title (C). In studies found that the linear kernel is the best kernel in classifying the web document compared to the RBF, polynomial and sigmoid kernel.

Youguang Chen, Jun Guo, Xue Deng and Min Zhu [8] in (2011) proposed an approach for document orientation detection and classification by using support vector machine (SVM) theorem. First, all the characters in a document image will be isolated and some valid ones are selected. Using the valid characters, the document image will be vectorized to a 32- dimensional vector by the feature extracting. By training lots of samples, an SVM classifier can be obtained, and then the orientation of unknown document images can be classified. Experimental results show the accuracy of the proposed method is considerably higher than Bray Curtis distance, even for some bad samples.

Mita K. Dalal and Mukesh A. Zaveri [9] in (2011) proposed an automatic text classification. An automatic text classification is a semi-supervised machine learning task that automatically assigns a given document to a set of pre-defined categories based on its textual content and extracted features. Automatic text classification has important applications in content management, contextual search, opinion mining, product review analysis, spam filtering and text sentiment mining. They explains the generic strategy for automatic text classification and surveys existing solutions to major issues such as dealing with unstructured text, handling large number of attributes and selecting a machine learning technique appropriate to the text-classification application.

K. Gayathri and A. Marimuthu [12] in (2012) proposed a text classification based on the feature selection and preprocessing by reducing the dimensionality of the feature vector and increase the classification accuracy. They studied the advantages and disadvantages of K-nearest neighbor (KNN) classification and Support Vector Machine (SVM) classification in performing their classification tasks. In their investigation, they found that the well-performing KNN classification approach may suffer from less accurate than the SVM classification. As per review study, they configure the situations and reasons for KNN classification to fail in performing effective and efficient classification tasks. It is the nature and properties of the conventional KNN algorithm that restrict the KNN classifier to perform well in certain problem domain such as computationally intensive, especially when the size of the training set grows large. To overcome the problem of conventional KNN they proposed the SVM machine learning techniques. The SVM classification models which full fill the problem solving requirement of specified domains, hence contribute to better classification effectiveness and efficiency. SVM can be used as a discriminative document classifier and has been shown to be more accurate than most other techniques for classification tasks.

Zhenqiang Xu, Pengwei Li, and Yunxia Wang [13] in (2012) proposed a new effective approach to optimize the SVM -DT classifier and presented the research on text categorization using SVM-DT classifier. The classifier of SVM decision tree (SVM-DT) takes advantage of both the efficient computation of the tree architecture and the high classification accuracy of SVMs. In this approach, a novel separability measure is defined base on Support vector domain description (SVDD), and an improved SVMDT is proposed. The basic idea of SVM-DT is that a class can be divided into the problem into a series of two-class problem, and the two-class problem can be solved by SVM. The

Support Vector Data Description (SVDD) is one of the methods which were used to describe data. A result shows the effectiveness and efficiency of the improved SVM decision tree.

With the increasing number of digital documents, the ability to automatically classify those documents both quickly and accurately is becoming more critical and difficult. Yan LI and Chungang CHEN [14] in (2012) developed a text classification system for Chinese documents. A HTF-WDF algorithm is proposed for feature selection. Different from other feature selection algorithms, this method considers the effect of term frequency. Using the idea of fuzzy feature, the terms with high term frequency (HTF) are distinguished and appended to the feature list. The features which can represent the topic of the documents are picked out according to the weighted document frequencies (WDF), which can avoid the problems of the traditional document frequency (DF) method. Then the Support Vector Machine (SVM) is used to training the classifier.

Mohammad Khabbaz, Keivan Kianmehr, and Reda Alhajj [15] in (2012) integrate soft clustering of words and feature reduction process for constructing set of informative feature vectors that represents both structural and textual aspects of XML documents. To extract structural information, they employ an existing frequent tree-mining algorithm combined with an information gain filter to retrieve the most informative substructures from XML documents. However, for extracting content information, they proposed soft clustering of words using each cluster as a textual feature. They conducted extensive experiments on a benchmark dataset, namely 20NewsGroups, and an XML documents dataset given in LOGML that describes the web-server logs of user sessions. With regards to the classifier built only using our textual features, the results show that it outperforms a naive support-vector-machine (SVM)-based classifier, as well as an information retrieval classifier (IRC). By applying SVM and decision tree algorithms using feature vector representation of XML documents dataset, achieved 85.79% and 87.04% classification accuracy, respectively, which are higher than accuracy achieved by XRules, a well-known structural-based XML document classifier.

Monika Arora, Uma Kanjilal and Dinesh Varshney [16] in (2012) developed a model for the efficient and intelligent retrieval. In this model they attempted to figure out the important factors for the successful efficient and intelligent retrieval. This model is designed to collate all the differing views on information retrieval to construct a holistic theoretical which is considered to be the source of a system.

Xu qihua and Geng shuai [17] in (2012) proposed a new fast learning algorithm for large-scale SVM under the condition of sample aliasing. The aliasing sample points which are not the same class are eliminated first and then the relative boundary vectors (RBVs) are computed. According to the algorithm, not only the RBV sample itself, but a near RBV sample whose distance to the RBV is smaller than a certain value will also be selected for SVM training in order to prevent the loss of some critical sample points for the optimal hyperplane. The most important fact is that the classification accuracy may be kept almost the same as that obtained when the large-scale sample set is used directly for training. The simulation results prove this fast learning algorithm very effective and can be used as a good practical approach for large-scale SVM training.

Inoshika Dilrukshi, Kasun Zoysa, and Amitha Caldera [18] in (2013) proposed a technique is to classify news into different groups so that the user could identify the most popular news group in a given country for a given time. The short messages were extracted from Twitter micro blog. Several active news groups were chosen to extract the short messages. Each short message was classified manually into 12 groups. These classified data were used to train the machine learning techniques. A word of each short message was considered as features and a feature vector was created using bag-of-words approach in order to create the instances. The data were trained using SVM (Support Vector Machine) machine learning techniques. The main reason of using SVM is, SVM supports high dimensional data. Current research is a high dimensional problem as a large number of features will be collected using short messages. Cross validation was done in order to avoid the biasness of data. The performance of the system will be the effectiveness of the system. Thus precision and recall values are calculated to measure the performance of the system.

3. CONCLUSION

The performance given by the SVM is higher if it involves large dataset for generalization of problem. The major strength of SVM is that the training of data is relatively easy. After the investigation of different techniques for feature selection and document classification, it has been concluded that though different techniques used for feature selection and document classification provide different level of accuracy but still there is a need to achieve high level of accuracy rate by reducing the computation time.

REFERENCES

1. Z. Zheng and R. Srihari, “Optimally combining positive and negative features for text categorization”, Workshop on Learning from Imbalanced Datasets II, ICML, Washington DC, 2003.
2. Anirban Dasgupta, Petros Drineas, Boulos Harb, Vanja Josifovski and Michael W. Mahoney, “Feature Selection Methods for Text Classification”, KDD’07, August 12–15, San Jose, California, USA, pp. 230-239,2007.
3. Dino Isa, Lam Hong Lee, V.P. Kallimani, and R. RajKumar, “Text Document Preprocessing with the Bayes Formula for Classification Using the Support Vector Machine”. IEEE Transactions on Knowledge and Data Engineering archive Volume 20 Issue 9, (September 2008). Page(s): 1264 –1272, IEEE-2008.

4. Janez Brank, Dunja Mladenic, Marko Grobelnik and Natasa Milic-Frayling, "Feature Selection for the Classification of Large Document Collections", Journal of Universal Computer Science, vol. 14, pp. 1562-1596, 2008.
5. Debnath Bhattacharyya, Poulami Das, Debashis Ganguly, Kheyali Mitra, Purnendu Das, Samir Kumar Bandyopadhyay, Tai-hoon Kim, "Unstructured Document Categorization: A Study", International Journal of Signal Processing, Image Processing and Pattern Recognition, Page(s): 1566 –1578, 2009.
6. Atika Mustafa, Ali Akbar, and Ahmer Sultan,"Knowledge Discovery using Text Mining: A Programmable Implementation on Information Extraction and Categorization", International Journal of Multimedia and Ubiquitous Engineering Vol. 4, No. 2, April, 2009.
7. Othman, M. S.,Yusuf, L. M and Salim J, "Features Discovery for Web Classification using Support Vector Machine", 2010 International Conference on Intelligent Computing and Cognitive Informatics, pp.36 – 40, IEEE-2010.
8. Youguang Chen, Jun Guo, Xue Deng, Min Zhu, "A Method for Detecting Document Orientation by Using SVM Classifier", pp.47 –50, IEEE-2011.
9. [9] Mita K. Dalal and Mukesh A. Zaveri, "Automatic Text Classification: A Technical Review", International Journal of Computer Applications, Volume 28– No.2, pp. 37-40, August 2011.
10. Shweta Mayor, Bhasker Pant, "Document Classification Using Support Vector Machine", International Journal of Engineering Science and Technology (IJEST), Vol. 4 No. pp.1741 – 1745, April 2012.
11. Ashis Pradhan, "Support Vector Machine -A Survey", International Journal of Emerging Technology and Advanced Engineering, Volume 2, pp.82 – 85, 2012.
12. K. Gayathri and A. Marimuthu, "Text Document Pre-Processing with the KNN for Classification Using the SVM", 7th International Conference on Intelligent Systems and Control (ISCO 2013), pp.453 – 457, IEEE-2012.
13. Zhenqiang Xu, Pengwei Li, Yunxia Wang, **"**Text Classifier Based on an Improved SVM Decision Tree**",** International Conference on Medical Physics and Biomedical Engineering, 2012. pp. 1986 –1991, IEEE-2012.
14. Yan LI and Chungang CHEN, "Research on the Feature Selection Techniques Used in Text Classification", 9th International Conference on Fuzzy Systems and Knowledge Discovery, (FSKD 2012), IEEE-2012.
15. Mohammad Khabbaz, Keivan Kianmehr, and Reda Alhajj, "Employing Structural and Textual Feature Extraction for Semistructured Document Classification", IEEE Transactions on Systems, MAN and CYBERNETICS—Part C: Applications and Reviews, VOL. 42, NO. 6, pp.1566 –1578, IEEE-2012.
16. Monika Arora, Uma Kanjilal, Dinesh Varshney, "Efficient and Intelligent Information Retrieval using Support Vector machine (SVM)", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, pp.39 –43, January 2012.
17. Xu qihua and Geng shuai, "A Fast SVM Classification Learning Algorithm Used to Large Training Set", 2012 International Conference on Intelligent System Design and Engineering Application, pp.15 –19, IEEE-2012.
18. Inoshika Dilrukshi, Kasun Zoysa, and Amitha Caldera, "Twitter News Classification Using SVM", The 8th International Conference on Computer Science & Education (ICCSE 2013) April 26-28, 2013. Colombo, Sri Lanka. pp. 287 – 291, IEEE-2013.