Spam Detection Using Clustering-Based SVM

Darshit Pandya
Department of Computer Engineering
Indus University
Gujarat, India, 382115
+91 7874605121
darshitpandya211@gmail.com

ABSTRACT

Spam detection task is of much more importance than earlier due to the increase in the use of messaging and mailing services. Efficient classification in such a variety of messages is a comparatively onerous task. There are a variety of machine learning algorithms used for spam detection, one of which is Support Vector Machine, also known as SVM. SVM is widely used to classify text-based documents. Though SVM is a widely used technique in document classification, its performance in the spam classification is not the best due to the uneven density of the training data. In order to improve the efficiency of SVM, I introduce a clustering-based SVM method. The training data is pre-processed using clustering algorithms and then the SVM classifier is implemented on the processed dataset. This method would increase the performance by overcoming the problem of uneven distribution of training data. The experimental results show that the performance is improved compared to that of SVM.

CCS Concepts

• Computing methodologies→Support vector machines
• Computing methodologies→Cluster analysis.

Keywords

Text Classification; SVM; Clustering.

1. INTRODUCTION

In the rapidly transforming world, the use of Email or SMS has been increased. With the increasing vogue of smartphones, the postal letters have been overshadowed by Email and SMS. Although technology has made our lives easy, there are some threats concerning the misuse of it. Due to the enormous use of these smartphone services, the threat of misuse is a huge concern. One of the common misuse of these services is Spamming. The attackers have been using this technique since the renaissance of the services with an intent to disturb the user. There are many reasons behind the spam attacks: Advertisement, Offers and many more. The spamming technique can also be used for criminal activities: spreading malicious software such as Computer Virus and Trojan. As the attackers have a gigantic target group, this activity is a major concern for the Cyber Security department of government. To solve the problem, many techniques have been developed for spam detection. These techniques were based on text categorization. As an advancement, some machine learning algorithms have also been given a try for spam detection. Some of the machine learning classification techniques used include Na ïve-Bayes Classifier[1], KNN[2], SVM, Decision Trees[3] and Random Forests[4]. Although some of these techniques have proved to be efficient, the efficiency achieved is somewhat low considering the amount of data.

In order to improve the efficiency of spam detection to a further extent, I propose a modified version of the primitive SVM classification technique. The proposed method works on text-classification as well as text-clustering methods. First, the system computes different clusters of the content of the messages using a simple k-means algorithm. Then, the outliers are removed from each cluster. Finally, the processed training set is provided to SVM for training a classifier model. As a new message arrives the content of the message is passed to the trained model, which categorizes the message as 'SPAM' or 'SAFE'. In this paper, I present my spam detection system and provide the results of my experiments using the SMS SPAM COLLECTION DATASET[5].

2. MOTIVATION

Majority of people using the Email or SMS services are victims of the spam attacks. Consequently, their privacy is at risk. Most of the users are either unable to distinguish spam messages or does not pay much attention to such messages. In both the cases, the attackers triumph. To protect the users from such a threat, there are many automatic spam detection systems available. However, their efficiency might not be that great so as to be considered reliable. In an attempt to improve the efficiency of the spam detection system, I present a more efficient system. My main aim was to provide a reliable method that could generate maximum performance possible and consequently, decrease the risk of spam attacks to a great extent. This method would help prevent spam attacks with a reliable efficiency and it would also be a trustworthy method for protection against the threats to privacy.

3. BACKGROUND

The method proposed under this paper is inspired from 'A Clustering-Based KNN Improved Algorithm CLKNN for Text Classification' by Lijuan Zhou, Linshuang Wang, Xuebin Ge and Qian Shi[6]. The idea in the mentioned paper describes the benefit of clustering the dataset before training the model. It gives a fair idea of improving the efficiency of the classification technique. They have used KNN as a classifier but in this paper, I have used SVM as the classifier. Another difference is that I have proposed a model that learns new data by combining it with the previous processed data and this combined dataset is further processed according to the proposed algorithm, which is then given as an input to the SVM and the predictions are made accordingly.

MLMI 2019, September 18–20, 2019, Jakarta, Indonesia
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-7248-0/19/09…$15.00
https://doi.org/10.1145/3366750.3366754

4. CONVENTIONAL SYSTEM

This system uses the SVM as a classifier. This method directly provides the unmodified original dataset to SVM. The classifier model is then trained and tested accordingly.

Algorithm: [**Classification using SVM**]

```
Input: SMS SPAM COLLECTION DATASET (D)
Output: Class label- ‘Spam’ or ‘Ham’
Step-1: Divide the dataset D into training (Tr) and testing (Ts)
        samples.
Step-2: Train the SVM model using Tr dataset
Step-3: Test the trained model using the Ts dataset
Step-4: Measure the accuracy
```

This is a very basic approach towards text categorization. It simply takes a training set as input, fits it using SVM and then predicts the output using the trained model. We need to improve the efficiency of this system so as to get a more reliable classifier.

5. PROPOSED SYSTEM

To improve the efficiency of existing SVM based spam detection system, I present a system based on clustering and classification techniques. The proposed system is divided into two algorithms-one for clustering and the other for classification. It is divided into two core parts: Pre-processing and Classification. Here the complete dataset is used in the pre-processing step. This logic greatly improves the efficiency of the system. Since this system is a combination of clustering and classification, I call it as CLUSTERING BASED SUPPORT VECTOR MACHINE or CLSVM. The steps involved in the proposed system are described in detail below.

5.1 The Pre-processing Step

This step includes the pre-processing of the dataset. Here, the whole dataset is pre-processed. Taking the full dataset while clustering is a unique feature of this system which helps to clean the data to a great extent and in turn, helps to improve the performance of the system as a whole. The idea behind clustering is to improve the quality of the unevenly distributed data. However, the processing time increases as the complete dataset is taken into consideration. But this drawback is overtaken by the improved efficiency of the output predictions. The pre-processing is achieved by following the below mentioned algorithm. Figure 1 shows the basic idea behind the pre-processing/clustering method.

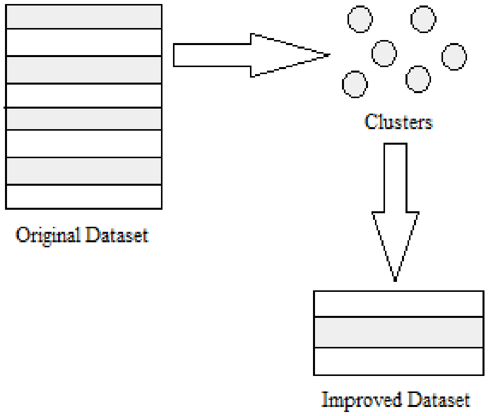

Figure 1. The pre-processing step.

Algorithm 1 [Clustering of the complete dataset]

```
Input: SMS SPAM COLLECTION DATASET (Train + Test),
       Number of samples N
Output: Improved Dataset
Step-1: Determine the number of clusters to be formed(X) using
        ‘The Elbow Method*’ from Fig. 2
Step-2: Convert text data to vectors using TF-IDF Vectorizer
Step-3: Form clusters using k means clustering
Step-4: For i : 1 to X do
          S :=  number of ‘Spam’ labelled text;
          H := number of  ‘Ham’ labelled text;
          If  len(S) > len(H) then
                  remove entries(non-test) with label ‘Ham’
          Else
                  remove entries(non-test) with label ‘Spam’
          EndIf
        Done
Step-5:  Return the remaining entries in the dataset
```

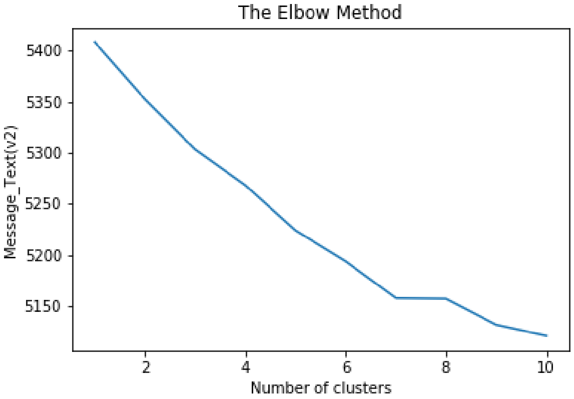

Figure 2. The elbow curve.

5.2 The Classification Step

After the pre-processing step, we have an improved dataset. Now, comes the main algorithm for classification of the improved dataset.

Algorithm 2 [Classification of the improved dataset using SVM]

```
Input: Improved dataset (D)
Output: Class label- ‘Spam’ or ‘Ham’
Step-1: Form training (Tr) and testing (Ts) samples.
Step-2: Train the SVM model using Tr dataset
Step-3: Test the trained model using the Ts dataset.
Step-4: Measure the accuracy
```

This step is similar to the conventional system mentioned before with the only difference being the quality of the dataset used as an input to the SVM classifier. Figure 3 shows the basic idea behind the classification step.

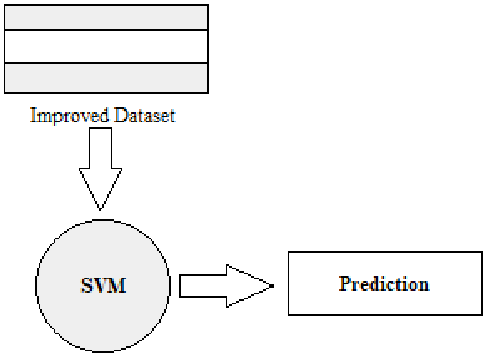

Figure 3. The classification step.

6. PERFORMANCE MEASURE

Now, let us compare the efficiency/accuracy of both- the conventional system as well as the proposed system. Here, the following resources are used for implementation:

Dataset: SMS SPAM COLLECTION DATASET [5] provided by the UCI Machine Learning Repository is used.

No. of Records: 5572

Model: Support Vector Machine

6.1 Conventional System

This system has an accuracy of 98.32435%. This performance is better as compared to other classification algorithms but it is not the best. The total time taken for execution was 32.39 seconds. The complete performance description is given in Table 1.

Table 1. Performance of conventional system

Total Records	Incorrectly Classified	Accuracy (%)	Time (s)
5572	94	98.32435	32.39

As seen from the table, the accuracy of the system is quite high. But if we consider the total number of samples, the accuracy may be considered a bit low as in a general scenario, the number of samples would be much more than the ones being tested in this dataset. Looking at the execution time for the complete system (Training + Testing), it is quite high compared to the number of observations taken into consideration. Suppose there are a 10 million records in the sample dataset. Now, according to the above observations, the net execution time required would be around 16 hours- which is quite a large amount of time and it would not be feasible to spend such a huge amount of time for a task that would not have a complete 100% accuracy. The data in the real-world applications would definitely be in millions and hence, this method of spam detection might not be feasible to use unless some sort of high end computing services are used for implementing the system.

6.2 Proposed System

As the proposed system is an improvement over the conventional system, its accuracy is also better than conventional system. The accuracy of the proposed system has been found out to be 100%. Now, this number is a great number in terms of reliability. The execution time for the proposed system was around 14.43 seconds, which is quite feasible considering the no. of records. The complete performance description is given in Table 2.

Table 2. Performance of proposed system

Total Records	Incorrect Classified	Accuracy (%)	Time (s)
5064	0	100	14.43

Here the total records are only 5064 instead of 5572 because the remaining records have been filtered out as outliers during the pre-processing phase of the proposed system. It can be clearly inferred from the Table 1 and Table 2 that the performance of CLSVM system is much better than the performance of the conventional SVM system in terms of both accuracy and execution time. Hence, the proposed system might serve a huge advantage for real-world applications where data is in enormous quantity and the execution time to process such data is less. Figure 4 visualizes the performance comparison between the conventional system and the proposed system. The relation between accuracy and execution time for both the systems clearly sheds light on the performance improvement achieved by the proposed system.

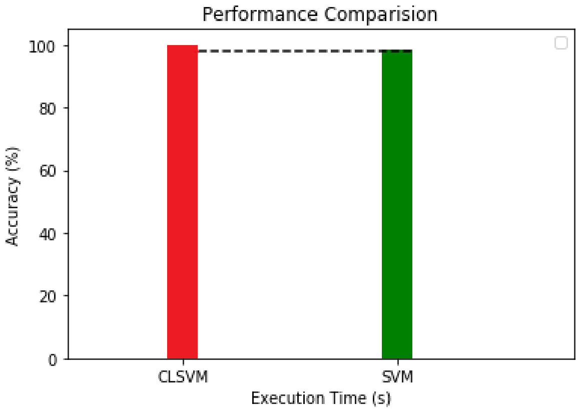

Figure 4. Performance comparison of SVM and CLSVM.

Although accuracy depends on the type of training and testing records, which are different for different datasets, the clustering technique, used to diminish outliers which in-turn improves the training set, forms an integral part for improving the accuracy of the dataset as a rich and uniform training set is presented to the classifier. The classifier accuracy largely depends on the quality of training set presented to it as an input. Consequently, improved training set improves the overall efficiency of the classifier used for classification of the dataset.

7. CONCLUSION

An improved SVM classification algorithm CLSVM which is an amalgamation of clustering and classification algorithms is introduced in this paper. In comparison with the conventional SVM classification technique, the new proposed system CLSVM uses K Means clustering technique in order to clean-up the outliers in the dataset and then the dataset is classified according to the SVM classification algorithm. This improved procedure has led to a great improvement in the accuracy rate as well as in the execution time of the system.

8. FUTURE WORK

Although the results obtained from the proposed method are great in terms of accuracy, there are few aspects which may have a flaw while working with a different dataset. One such flaw could be the case when the Elbow Curve would become smooth for a dataset which is not very clustered. Such a challenge can be addressed by using some other alternative methods such as The Gap Statistic method [7], which would be carried out in continuation of this work. Also, the proposed system in this work focuses on a well-known dataset. However, the work can be extended by using other well-known datasets which in-turn might help detect any flaws in the proposed system. Also, a different clustering algorithm can be implemented instead of K Means which may provide better accuracy as compared to the traditional K Means clustering.

9. ACKNOWLEDGMENTS

My thanks to ACM SIGCHI for allowing us to modify templates they had developed. My sincere thanks to UCI (Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science) for letting me use it’s Machine Learning Repository on the basis of which the experimental results mentioned in the paper are calculated. My sincere thanks to the professors of computer engineering department, Indus University for encouraging and supporting my work. I would also like to thank Mr.Varun Gohil for his constant support and encouragement.

10. REFERENCES

[1] Jeremy Eberhardt. Bayesian Spam Detection. 2014. *UMM CSci Senior Seminar Conference,* December 2014 Morris, MN.

[2] S.Ananthi, Dr. S. Sathyabama. 2009. Spam Filtering Using K-NN. *Journal of Computer Applications,* Vol-II, No.3, July-Sep 2009, 20-23.

[3] Sarit Chakraborty, Bikromadittya Mondal. 2012. Spam Mail Filtering Technique using Different Decision Tree Classifiers through Data Mining Approach - A Comparative Performance Analysis. *International Journal of Computer Applications,* Volume 47– No.16, June 2012, 0975 – 888, DOI= 10.5120/7274-0435

[4] Manish Kumar. 2016. Effective Spam Filtering using Random Forest Machine Learning Algorithm. *International Journal of Innovative Research in Computer and Communication Engineering.* Vol. 4, Issue 3, March 2016, 3200-3205 DOI= 10.15680/IJIRCCE.2016. 0403048

[5] SMS Spam Collection Data Set, UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection

[6] Lijuan Zhou, Linshuang Wang, Xuebin Ge, & Qian Shi. 2010. A clustering-Based KNN improved algorithm CLKNN for text classification. *2010 2nd International Asia Conference on Informatics in Control, Automation and Robotics (CAR 2010).* DOI= 10.1109/car.2010.5456668

[7] Tibshirani R, Walther G, Hastie T. 2001. Estimating the Number of Clusters in a Data Set via the Gap Statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology*). 63(2):411-423. DOI= 10.1111/1467-9868.00293

[8] Lee, J., & Lee, J.-H. (2014). K-means clustering based SVM ensemble methods for imbalanced data problem. 2014 Joint *7th International Conference on Soft Computing and Intelligent Systems (SCIS) and 15th International Symposium on Advanced Intelligent Systems (ISIS).* DOI= 10.1109/scis-isis.2014.7044861

[9] Sasaki, M., & Shinnou, H. 2005. *Spam detection using text clustering. 2005 International Conference on Cyberworlds (CW’05).* DOI= 10.1109/cw.2005.83

[10] Sculley, D., & Wachman, G. M. 2007. Relaxed online SVMs for spam filtering. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR ’07.*

DOI= 10.1145/1277741.1277813

[11] Lee, L. H., Wan, C. H., Rajkumar, R., & Isa, D. 2011. An enhanced Support Vector Machine classification framework by using Euclidean distance function for text document categorization. *Applied Intelligence*, 37(1), 80–99. DOI= 10.1007/s10489-011-0314-z

[12] ZHANG Yun-tao, GONG Ling, WANG Yong-cheng. 2005. An improved TF-IDF approach for text classification. *Journal of Zhejiang University SCIENCE.* August 2005, Volume 6, Issue 1 DOI= 10.1631/jzus.2005.A0049

[13] Anubhav Aggarwal, Jasmeet Singh, Dr. Kapil Gupta. 2018. A Review of Different Text Categorization Techniques. *International Journal of Engineering & Technology*, 7 (3.8) 2018, 11-15. DOI= 10.14419/ijet.v7i3.8.15210

[14] Wang, J., & Su, X. 2011. An improved K-Means clustering algorithm. *2011 IEEE 3rd International Conference on Communication Software and Networks*. DOI= 10.1109/iccsn.2011.6014384

[15] Dora Arul Selvi Balasingh, Kamaraj Nagappan. 2009. Support Vector Classifier with Enhanced Feature Selection for Transient Stability Evaluation. *SERBIAN JOURNAL OF ELECTRICAL ENGINEERING,* Vol. 6, No. 1, May 2009. UDK= 514.742.2:621.31