Hybrid Spam Detection Model

Meghamala Sinha	Purbasha Chatterjee	Alex C. Way
Oregon State University	Oregon State University	Oregon State University
121 NW 21st St Corvallis OR	121 NW 21st St Corvallis OR	4035 Scenic Dr. Eugene OR
sinham@oregonstate.edu	*chattepu@oregonstate.edu*	*waya@oregonstate.edu*

Abstract

Unwanted emails (spam) has been an increasing problem since the 1990's to the internet community. The number of span messages is huge, and the total number of emails to check for spam is even larger. More efficient and effective algorithms for detecting spam become necessary as the amount of spam increases. We review and compare the efficiency and effectiveness of four contemporary algorithms (SVM, Neural Networks, Decision Trees, and Random Forest). We also propose a new approach to merge the essence of clustering (K-means) and classification algorithms for training on our dataset. The results show that our approach hugely reduces the training time, improves accuracy by reducing the scalability of the data and thereby offering a more generalized performance.

1 Introduction

Electronic mail is one of the cheapest and easiest means of communication [1]. It has helped billions of people to get connected globally, enabling us to send several e-mails to different people within a short span of time. It not only supports conversation, but also acts as a task manager, document delivery system, and an archive. With this increasing demand of e-mail communication, many fraudulent activities have also erupted, disrupting the security of data.

A variety of technical measures have been proposed. One of the effective solutions is to detect spam messages automatically by building a model to detect spam. These models are often called spam filtering or spam detection models. Spam detection models are divided into two primary types of approaches: non-statistical and statistical approaches. The latter is considered to be more powerful than the former. Many of the statistical detection models that exist, search for particular keyword patterns in the messages. Several spam detection models using machine learning techniques have been researched.

Different types of Machine Learning methods have been proposed to classify emails to spam or not-spam [2]. These methods have the power to extract and learn important information out of the E-mails and use these gained knowledges to categorize a new incoming email and identify it into either spam or not spam. There are various types of machine learning methods namely supervised learning, unsupervised learning and semi-supervised learning.

For the development of spam-filters it has been seen that mainly supervised leaning method have been broadly studied and implemented. Supervised learning [3] is used for classification of data. It mainly involves two stages- 1) the leaning or training stage, 2) prediction based on established knowledge. Unsupervised learning [4] is mainly used for grouping data into clusters based on some similarity measures and finding relationship between them.

2 Objective

In this project, we want to present a hybrid approach of classification over clustering. We used K-means clustering (since it is the strongest clustering method) to derive an optimal dataset. This optimal dataset is used to train several classification models. Results showed effective decrease in training time and increase in prediction performance. The motivation of this project is to demonstrate a comparison between the performance of optimal dataset over the original dataset by SVM, Neural Network, decision tree, and random forest models.

3 Approach

In this section, we present our proposed work to meet the objective of our project. Several classification methods like SVM, Random Forest, Neural Network have been used to compare the results with our proposed model. Support Vector Machine [5] is a discriminative classifier which provides a representation of data as points in space which are mapped in such a way that different examples are separated by a clear gap. New data points can be mapped on the space and predicted to a specific category depending on which side they are placed. Decision tree aims to create a model that predicts the value of a target variable based on several input variables. It classifies the training set based on labelling of each internal node with an input feature. On the other hand, Random Forests [6] is a type of ensemble learning techniques which concerns at the noise and the number of attributes. It builds an ensemble of classified trees using bagging mechanism. It runs efficiently on large data sets with many features with fast execution speed. Apart from these, when a neural network model [7] is trained, it means selecting one model from the set of allowed models that minimizes the cost criterion. Several algorithms have been developed for training neural network models which mainly focuses at optimization theory and statistical estimation. K-means [8] is a clustering algorithm which aims to divide data into groups based on similarity. Similarity is an important role in this type of unsupervised learning. Similarity is measured using Euclidean Distance. Steps involve in K-means are as follows-

1)At the beginning we randomly choose K points as the initial cluster center.
2) *REPEAT*
3) Each data point is assigned to the most similar cluster based on the mean value of all the objects in a cluster.
4) We update the mean value of a cluster
5) *UNTIL* the mean values of clusters not change.

We started our implementation by first creating a dictionary of unique words from our dataset. This dictionary was used to extract the frequency of each word in each mail to create the features of the dataset. Then several classification strategies were applied to this newly created dataset.
After that we have used stemming and stop word removal was used for data preprocessing over the dataset.

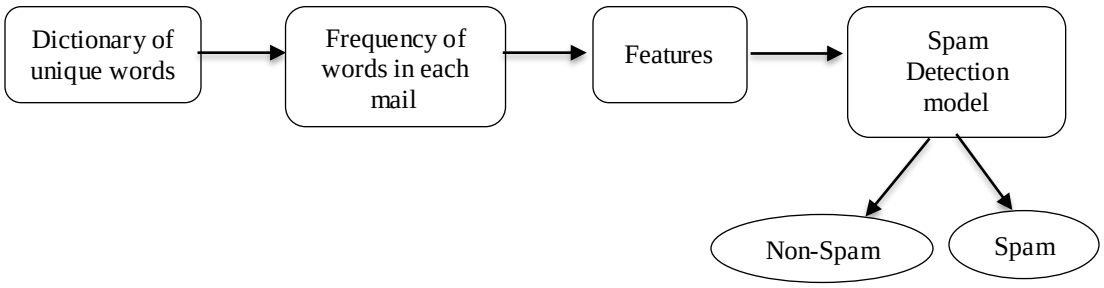


Figure 1: Flow of the model

We have used MATLAB machine learning toolbox to implement the classification. In case of neural network, we divided the data randomly into training (70% of data), validation (15% of data), and testing (15% of data) sets. This helped us to provide variance in the training data. Currently, we could run the training data once, but we could run it in multiple random successions to potentially increase accuracy. We use 10 hidden neurons in our pattern recognition network. We also use cross-entropy in our loss calculation. In case of SVM, decision tree and random forest, we applied k-fold cross validation to extract the accuracy.

We progressed with unsupervised learning technique by performing K-means on training set so that each sample will get a cluster label. The main task of clustering is to group the objects into clusters, objects in the same cluster are of similar pattern than those in different clusters. The selected clustering result have massive information and important contribution for building the classification model. We trained over selected training set which we have achieved through K-means clustering and then applied SVM/ Random Forest and Neural network classification over it to get the overall accuracy and computational cost.

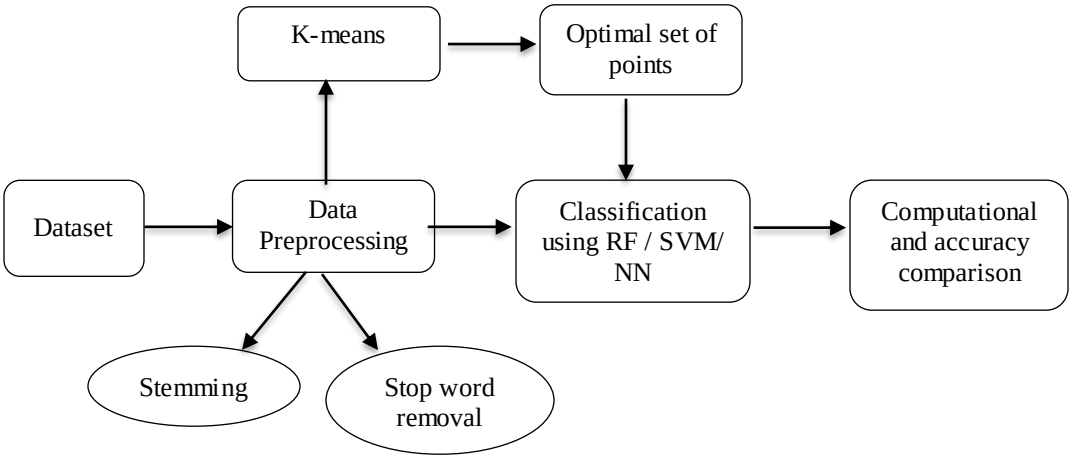


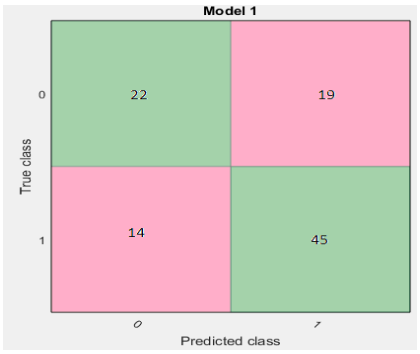
Figure 2: Workflow of our algorithm

4 Experiments/Results

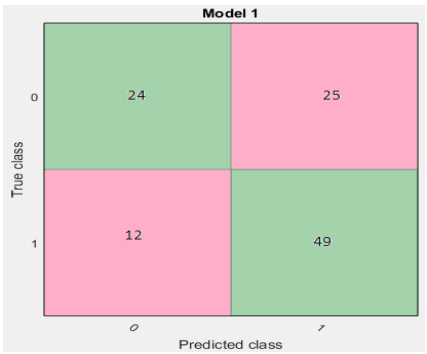
We use the database from "http://csmining.org/index.php/spam-email-datasets-.html" to test the algorithms.

Model	Accuracy (%)
Decision Tree	63.44
Neural Network	64.02
SVM	66.67
Random Forest	75.12

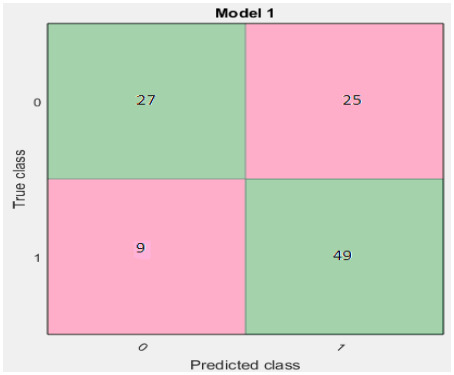
Table I: The Accuracy rate



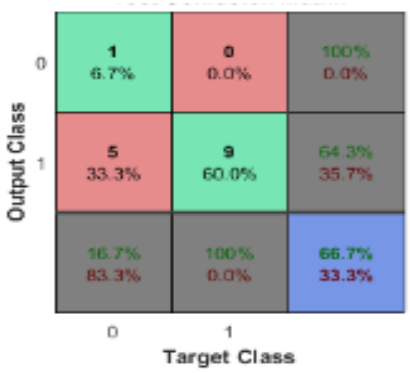
SVM



Decision tree



Random forest



Neural Network

Figure 3: Confusion matrix of all classifications

As we can see from our Table I and Figure 3 that the accuracy rate of all the classifications are overall not high. Most of the accuracy are in the range of 60%. Hence, we applied these classifications over clustered data with an expectation to have an improvement in accuracy and reduce the computational cost.

Model	Accuracy (without K-means)	Accuracy (with K-means)	Accuracy Difference
SVM	66.67	71.43	4.76
Random Forest	75.12	76.19	1.07
Neural Network	64.01	64.02	0.001

Table II: The comparison of the accuracy rate

After applying classification over clustered data, we achieved a great improvement in accuracy in case of SVM. The accuracy rate went high by almost 5% as you can see in Table II. In case of neural network, the accuracy rate remains almost same but we achieve a great decrement in computational cost as tabulated in Table III.

Model	Time without K-means (sec)	Time with K-means (sec)
Neural Network	8.635	1.557
Random Forest	5.201	3.56
SVM	1.125	0.6935

Table III: The comparison of the computational cost

We analyze the overall accuracy increment and computational decrement as shown in Figure 4 and Figure 5.

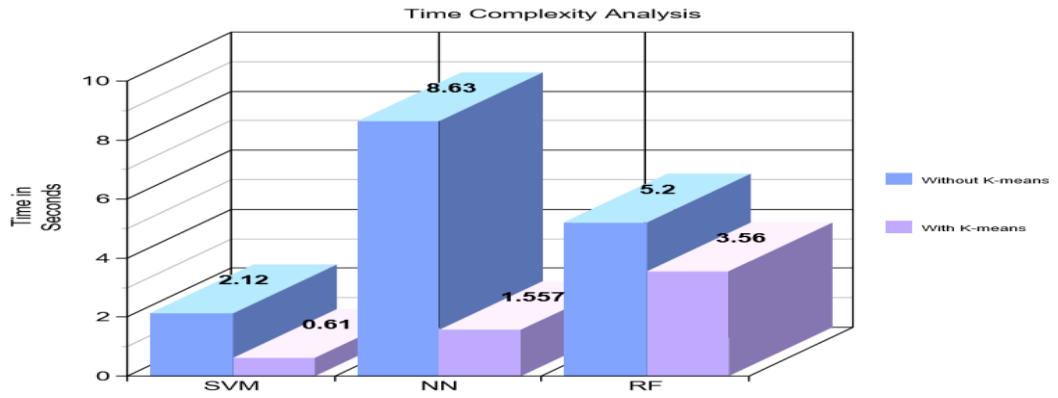


Figure 4: Analysis of time complexity

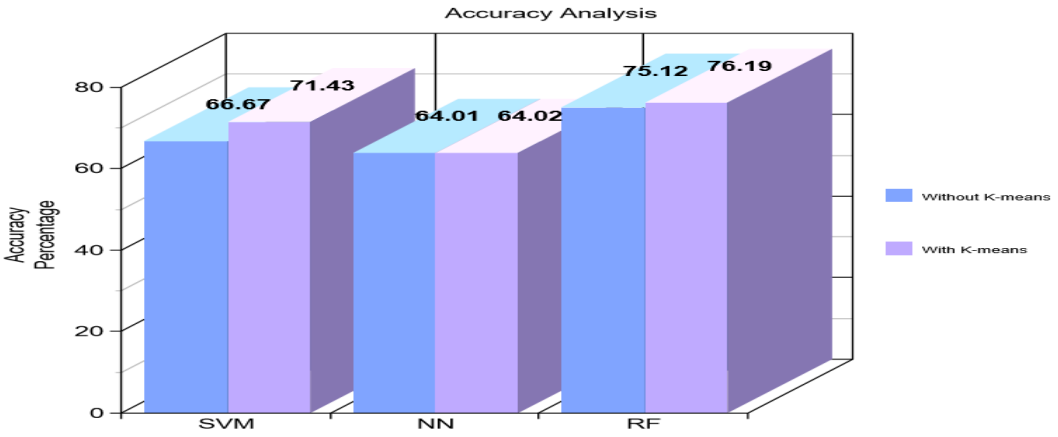


Figure 5: Analysis of accuracy rate

5 Conclusion

In this project, we propose a new hybrid method of spam detection by combining the benefits of classification over clustering. The motivation behind using K-means clustering algorithm is to group the messages or emails based on some features or similarity into disjoint clusters. This reduces the training time as well as increased the accuracy. Specially in the case of Neural Network, the training time is reduced by 82%. Less amount of work

has been done using this approach.

Novelty of our work- There has been some past work dome to reduce computational cost using K means + SVM [9]. But in this project, we have also improved the prediction accuracy. Furthermore, we have also shown huge reduction in the computational training time for K-Means + NN in this object. We can claim that this type of hybrid model is a novel approach as no work has been yet done in this field.

Future Work- We can enhance our model with more advanced featurization to our dataset to observe similar significance in the accuracy as well. Some of the future aspects of our project is as follows:

- Use more number of dataset to get better accuracy result.
- Explore better ways to select support vectors
- Explore for more featurization

References

[1] Caruana, G. and M. Li, 2012. A survey of emerging approaches to spam filtering. ACM Comput. Surv. (CSUR), 44(2): 9.

[2] Tretyakov, Konstantin. "Machine learning techniques in spam filtering." *Data Mining Problem-oriented Seminar, MTAT*. Vol. 3. No. 177. 2004.

[3] Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas, 2007. Supervised machine learning: A review of classification techniques.: 3-24.

[4] Yu Zong, Ping Jin, Dongguan Xu, Rong Pan, 2013. "A Clustering Algorithm based on Local Accumulative Knowledge", Journal of Computers, pp.365-371, vol.8, no.2.

[5] Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin. "A practical guide to support vector classification." (2003): 1-16.

[6] Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." *R news* 2.3 (2002): 18-22.

[7] Wang, Sun-Chong. "Artificial neural network." *Interdisciplinary Computing in Java Programming*. Springer US, 2003. 81-100.

[8] Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1 (1979): 100-108.

[9] Yao, Yukai, et al. "K-SVM: An effective SVM algorithm based on K-means clustering." *Journal of computers* 8.10 (2013): 2632-2639.