Content Based Classification using Clustering Algorithm-Survey Paper
Patel Vaishali1 Dave Nakul2 Dave Avani3
1Department of Computer Science & Engineering 2,3Department of Computer Engineering
1SRI, Vadasama, Kalol2,3VGEC, Chandkheda, Ahmedabad

Abstract— This survey paper categorizes, compares, and summarizes from some published technical and review articles in automated fraud detection within the last few year. Short Message Service (SMS) has become one of the most common communication method due to fast increment in the number of users worldwide. As usage of SMS increased spam messages also increased. To filter spam SMS there are many methods available like Support vector machine(SVM), Naive Bayesian, K-nearest neighbor(KNN), K-star, Feature extraction, Feature selection, Tokenizer etc. Naive Bayesian is considered as one of the most effectual and significant learning algorithm for machine learning and data mining and also has been treated as a core technique in information retrieval. Feature extraction is the best method for SMS spam filtering. K-nearest neighbor algorithm is simple to understand but non parametric lazy learning algorithm. Support vector machine is the feature based classifier and it has high accuracy.

Key words: SMS, support vector machine, Naive Bayesian, k-nearest neighbor, feature extraction

I. INTRODUCTION

SMS (Short Message Service) is the simplest and most common type of text message and is supported by a large number of mobile phones. SMS typically allows maximum of 160 characters per message. In recent years, Short Message Service (SMS) has one of the most common method in all other communication method due to rapid increase in the number of mobile phone users around world. This increase has unavoidably attracted spammers and caused SMS spam (unsolicited) message just as in the case of spam e-mails. Today, majority of SMS messages received by mobile phones are unfortunately disturbing spam messages such as credit opportunities of banks, promotion and discount announcements of stores, and new tariffs of communications service providers. According to the regular usage of SMS, spam-filtering systems are becoming more widespread. Spam filtering systems added on cell phones a defined number of specific words, and this software classifies SMS according to these keywords. Yet, this technique cannot correctly filter each kind of SMS. Though the problem of SMS spam is not as old as of email spam there have been several efforts in the literature to detect SMS spam messages. Some examples to those efforts are as follows. Bayesian filtering techniques were employed. Feature-based and compression model-based filters were evaluated. Bayesian learning, support vector machine classification, k-nearest neighbor and k-star were used in [1]. Bayesian based classifiers were utilized together with the distinctive features determined by information theoretic feature selection methods. Finally, a number of recent studies on SMS spam filtering are reviewed. This work mainly focus on how to categorize spam and non-spam SMS which arrives to our inbox of our mobile phones.

II. NAIVE BAYESIAN CLASSIFIER

Naive Bayesian is simple and robust classification method. It is a method that depends on probabilistic conditions [2]. It uses Bayesian theorem that assumes between attributes and calculates the probability of them by computing the frequency of values and the relationship between them in data. It is good when the dimensionality of input is high and generally good classifier for text documents as well as where SMS spam filtering is considered text. It only assumes that attributes of the data set are conditionally independent while in reality this condition hardly holds. In the Naive Bayes classification, all words in a given SMS are considered as mutually independent. It is the simplest form of Bayesian network which can be interpreted as conditional independent [3]. The weakness of this classification method is that the assumption of class conditional dependencies does not hold and also can't modeled by this method.

III. K-NEAREST NEIGHBOR CLASSIFIER (KNN)

KNN is simple to implement. It is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure [4]. KNN computes the similarity between the new data and the training data set stored previously. Then, according to the predefined value, K, the most K similar data are listed in descending order. Finally, the new data takes the class label that belongs to the majority neighbors. It is not good classifier for text documents and SMS spam filtering. KNN requires the selection of a single parameter K. KNN suffers from a random selection of K value that may end up with miserable outcomes, and also it is sensitive to irrelevant parameters such as determining the proper similarity measure [2]. The weakness of KNN is, its need a lot of space to store all data and takes more time to classify a new data than with a model. The disadvantages of this method are memory intensive and classification are slow.

IV. SUPPORT VECTOR MACHINE CLASSIFIER (SVM)

SVM method is used to classify the n-dimensional hyper plane into only two classes. SVM are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. SVM support both classification and regression task. It is best classification method for text classification and sms filtering [2]. SVM has high accuracy so many researcher widely use this classification method for text classification. Complexity is the main weakness of this method.

V. FEATURE EXTRACTION

Feature extraction involves reducing the amount of resources required to describe a large set of data. In this
method bag-of-word model is used. In that extract ordering of words are done. In which some terms are ignored but frequency or occurrence of that term is considered. This terms are assigned particular weights representing their importance in a given document. It is used for SMS spam filtering can be treated as conventional text classification task, the structure of spam messages can be significantly different than that of formal texts. The structural features (SF) extracted from a given SMS message into spam and ham [1]. This method is scalable and efficient so this is highly demanded to deal with high-dimensional document feature sets [5].

VI. EVALUATION

In this section, we provide detail about dataset, algorithm used in our comparison and analysis.

A. Data Set:
The dataset in our experiments is obtained from corpus [3]. In the corpus collection of spam sms is built from the public website and the messages are collected for mobile phone spam studies. The sms spam collection dataset are divide in two classes spam and ham.

B. Algorithm:
Due to the text formatting property of sms, we applied text to the dataset. After feature extraction are applied on text sms so text sms are converted into number of token. After all the classifier method are applied on that text. The classifier method applied on that are naive Bayesian, KNN and SVM.

C. Comparison and Analysis:
According to research, the classification method are applied on text the accuracy of all method are below [2].

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN</td>
<td>0.899</td>
<td>88.54%</td>
</tr>
<tr>
<td>NB</td>
<td>0.975</td>
<td>97.52%</td>
</tr>
<tr>
<td>SVM</td>
<td>0.986</td>
<td>98.6%</td>
</tr>
</tbody>
</table>

Table 1: Accuracy Table

According to above accuracy the SVM method is the best classification method whereas NB is a comparatively good than KNN. But NB gives less accuracy than SVM.

VII. CONCLUSION

In this paper some of classification and clustering method are described which is used for text classification and SMS filtration. Naive Bayesian is simple and good classifier when dimensionality input is high. KNN method is easy to implement but very slow classification. SVM has high accuracy and it is the best classification method for SMS filtering. Feature extraction method is used for to count the weights of term which are used in text document.

REFERENCES