Clustered K-Nearest Neighbor Process for Spam Detection in Social Networks

Jyotika Verma
PG Student
Department of Software Engineering
University Institute of Engineering and Technology
Kurukshetra University, 136119 Kurukshetra, INDIA

Dr. Sanjeev Dhawan
Assistant Professor
Department of Software Engineering
University Institute of Engineering and Technology
Kurukshetra University, 136119 Kurukshetra, INDIA

Abstract

Spam is one of the issues of the Internet which required attention the most especially on the social networking websites and that's why spam detection is a very concerned issue. Although there are various issues related to the Internet but spam is the one which is faced by everyone who is using the Internet. There are various strategies which are used by the researchers to detect the spam; this is discussed in the Literature Review section of this paper. KNN is one of those techniques and it is abbreviated as K-Nearest Neighbor Classification, it is basically used to find the nearest neighbor based on the pattern recognition and it is a non-parametric technique. Here, in this research paper KNN is used with the Clustering process because of which the process is time efficient and that too with great accuracy. The data which has been provided for the entire process to be performed is extracted from the social networking website Twitter with the help of R package as it provides interface with the Twitter web API (Application Programmable Interface). Various Calculations has been performed to calculate the accuracy, precision, and other parameters and based on these results and respective graphs have been obtained.

Keywords: Clustering, KNN, Preprocessing, R package, Social Networks, Spam, Stop word removal

I. INTRODUCTION

Internet is commonly used by every person as it is a boon which was given by Vint Cerf who is one of the fathers of Internet. There are various advantages of this technology but along with this disadvantages also came. Internet provides free access to everyone including children and old age persons too which is not a good thing. There are certain things on Internet which are not suitable for children to read or to watch or even to click on. To avoid this kind of issue "Parental Control Software" is used. This is one of the issues which are related to the Internet but the problem on which this paper is written about is "Spam". Every body knows about this spam. It is defined as the unwanted mails or information or anything which create hindrance in performing your work is a spam. It is very much seen in the inbox of one's email account but now this is very popular on Social Networking Websites. To cope up with this problem various techniques have been used which are called as spam detection techniques. Some are Supervised and others are Unsupervised, the only difference between them is that supervised techniques uses the trained data whereas unsupervised techniques uses the untrained data. The technique which is discussed in this paper is K-Nearest Neighbor technique which is a supervised one means it uses the trained data. KNN works well in every aspect except one and that is if KNN is to be used with large dataset then it will take a long time for processing so to handle this problem, a clustering approach which is a unsupervised technique is used along with KNN with the help of which different groups has been maintained which contains the data more similar to each other rather than the data present in other group and these groups are called as clusters [9]. Hence, the complete process is said to be as Clustered KNN. The data which has been provided as input to this process is extracted from Twitter with the help of R package. R package is a tool which is very much used for the statistical computing and graphics [10]. On the data which is extracted by R package further steps are performed including Preprocessing, Feature Extraction and then evaluation. Preprocessing is further performed in two steps i.e., stop word removal and stemming. At the end of this paper the results along with the conclusion and future scope are given and a system which is time efficient and accurate is introduced.

II. LITERATURE REVIEW

The present literature survey concentrates on the work done by a number of researchers Worldwide in the field of spam detection both in e-mail systems and social networking Websites. This literature survey is carried out in chronological order and is presented from year 1998 to 2014 as under. Several of work has been performed by *Naïve Bayes* for filtering the spam out of ham which was basically used for the text classification [1]. The first program which has used the Bayes classifier was of Jason Rennie's *ifile* program which was introduced in 1996[1]. The very first publication about the Bayesian work was done by Sahami *et al.* in 1998 [2]. Their method made it possible that the spam filters automatically learn to separate the junk email out of the legitimate emails and they have also improved the efficiency of the filters by providing the domain specific feature they the

filters didn't have to read the whole text but with the help of certain words the mail could be classified as spam or ham. But any of the solution used for the detection of the spam cannot solve the problem completely. The usage of the Internet is growing rapidly and with this the usage of the email is also increasing. Spam is referred to as Unsolicited Commercial Email (UCE). Each and every word in this sentence has its own meaning. The "unsolicited" means the receiver have not requested for that particular thing or the receiver have not asked for [3]. They have referred the spam as one of the easiest and cheapest way to communicate. Due to which the traffic over the Internet have also increased to an extent that sometimes the important mail didn't reach to the receiver and sometimes this have caused a major loss to user. To overcome this problem a novel-spam technique was presented by Yang Li *et al.* [4] which was developed on the basis of the user's feedback and the improved Bayes approach which results a good solution for filtering of the spam and have given the relatively lower false positives. All of the spam problems cannot be solved by using only one technique, because every technique has some problem in detection of the spam or ham. Spam has caused very much problem in our day to day life as it doesn't allow us to do our work efficiently. The spam always fills up the user's account inbox and because of this the legitimate emails doesn't get the chance to get addressed because sometimes with the purpose of getting this spam email out of the inbox we by chance delete the legitimate mails too. And the second problem related to the spam is sometimes there is a case when we encounter a large amount of spam in our email account and hence it takes a lot of time to get rid of those spam messages. There also chances that spam messages may contain the malicious content which may crash the user's system. It is experienced that the situation in China and America is worse that 50% of the mail which have been received or more than that is spam [5]. Federal Trade Commission is U.S commission which works for enhancing the security among the users and various economies. It address the various privacy issues including spam, social networking, peer to peer, mobile etc and all these matters include over 130 spam and spyware cases [6].Spam filter is defined as a program which is used for detecting the unsolicited email and prevents those messages from entering into the user's inbox [7]. This has been experienced that there is not any of the filter which gives the 100% result in identifying the spam or ham (the legitimate emails). Several of the anti-spam techniques have been used but only one technique cannot be chosen to perform for all of the work. Every technique is having its limitations beyond which it cannot perform well. Many of the techniques have been used for the detection of the spam email like KNN algorithm, AdaBoost, Chi-Square etc. But all of them were having some of the limitations then an Alliance based anti spam approach was introduced by Yu-Fen Chiu *et al.* [8]. This approach collects the interesting information related to the spam and then classifies, discover and exchange it. The spam filter which is proposed in here is based on the relative set theory, genetic algorithm and XCS classifier system. This was the research work which have been studied in the context of the spam detection and on the basis of the all the information it is concluded that the number of problems have been solved with all the techniques and number of algorithms are available for the spam detection.

III. IMPLEMENTATION

The work is performed in five steps namely; extracting the data, Preprocessing, Feature Extraction, Classifier, and Evaluation. Extraction involves the use of a R package named tool which helps to get the data from the Social Networking Website Twitter, Preprocessing is the method which is to be performed on the data provided so that the complexity can be reduced. It is performed in further two steps involving stop word removal and stemming. After this Feature Extraction is performed and for this certain parameters has been set and according to those parameters whole data gets checked. Later, clustering is performed and then the classifier is applied to the clusters and then evaluation is performed.

A. Extracting the data:

R package is the tool which is used to extract the data from Twitter. This tool is basically used for statistical Computing and graphical display. It is free environment software. It runs on windows, Linux and Mac OS. R can be easily extended with 6,600+ packages available on CRAN. Data in R can be saved as .Rdata files with function save() and then the loading of the data can be performed with the help of the load() function.

```
> a < -1:10
> save(a, file="./data/dumData.Rdata")
> rm(a)
> load("./data/dumData.Rdata")
> print(a)
[1] 1 2 3 4 5 6 7 8 9 10
```

Here, rm() function removes an object a from R.

R package in this research work has been used with the purpose of extracting the tweets from the Twitter Website. R package is an interface which is used with the Twitter Web API.

B. Preprocessing:

This is performed to reduce the complexity of the data provided by making it simple and easy to read. To do this two techniques have been used and these are:

1) Stop Word Removal:

This is the process of removing the unwanted and unnecessary words from the text. For this, a separate file is maintained with the name stop words, so that whenever those words comes in the input they automatically gets removed [11].

For example: if we have maintained a file and in that file the words written are 'a' an', the' etc. So, whenever these words will appear in the input they automatically get removed and hence, complexity is reduced of the document

2) Stemming:

This is the technique to find out the root or the stem of the word or in other way it is defined as the process of finding out the origin of the word [12].

For example: Relation, Relationships, Relative, Related are the words which has some origin and that is Relate.

C. Feature Extraction:

This is the process in which number of features set is maintained on the basis of which inpurt is examined. In this work , the set maintained is combination of five features and that is no. of spam words per input (scount), total number of words in a line (wordcount), no. of URL (url) , no. of URL present per wordcount (URL/wordcount), Retweets (rt).

For example: Input: RT @BBCSport: Stuart McCall appointed #Rangers manager until the end of the season http://t.co/oJRIISNwYI http://t.co/J1AXlQGfvP

This value is given as an input to the process and if it is mentioned that more than two links in one input is not allowed then that input is considered to be as spam. So, here because of the presence of the two hyperlinks "http://t.co/oJRIISNwYI, http://t.co/J1AXlQGfvP " this is considered to be as spam.

D. Classifier:

KNN classification is used which is basically a non-parametric method and used for classification and regression. It depends on the output of the system whether KNN is used for the classification or for regression.

The K-Nearest Neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its *k* nearest neighbors (*k* is a positive integer, typically small).

E. Evaluation:

On the basis of the process performed above, sensitivity, specificity/Recall, Precision , Accuracy are evaluated and then the corresponding graphs get generated.

IV. RESULTS AND ANALYSIS

Entire work is performed in MATLAB which is abbreviated as Matrix Laboratory, it is used to perform the data visualization, data analysis and it is a very high-level interactive programming language [13]. Clustered kNN is implemented in MATLAB for comparison of time efficiency and various other factors. For performing this work, a dataset of 150 records has been maintained and those records are basically the tweets, which are downloaded from Twitter using the R package tool and that data is saved as tdata.csv , here extension .csv means "comma separated values" A Comma Separated Values files stores tabular data in plain text. Plain text means that the file is interpreted a sequence of characters, so that it is human readable with a standard text editor. Each line of the file is a data record. On various Percentage of Trained data the accuracy, precision, recall has been analyzed.

Table - 1Results on the basis of Trained Data

Percentage/Parameter	Accuracy	Sensitivity	Specificity	Precision	Recall	Time
20 %	92	96	85	92	96	0.363
40%	93	100	70	90	100	0.019
80%	98	99	96	98	99	0.011

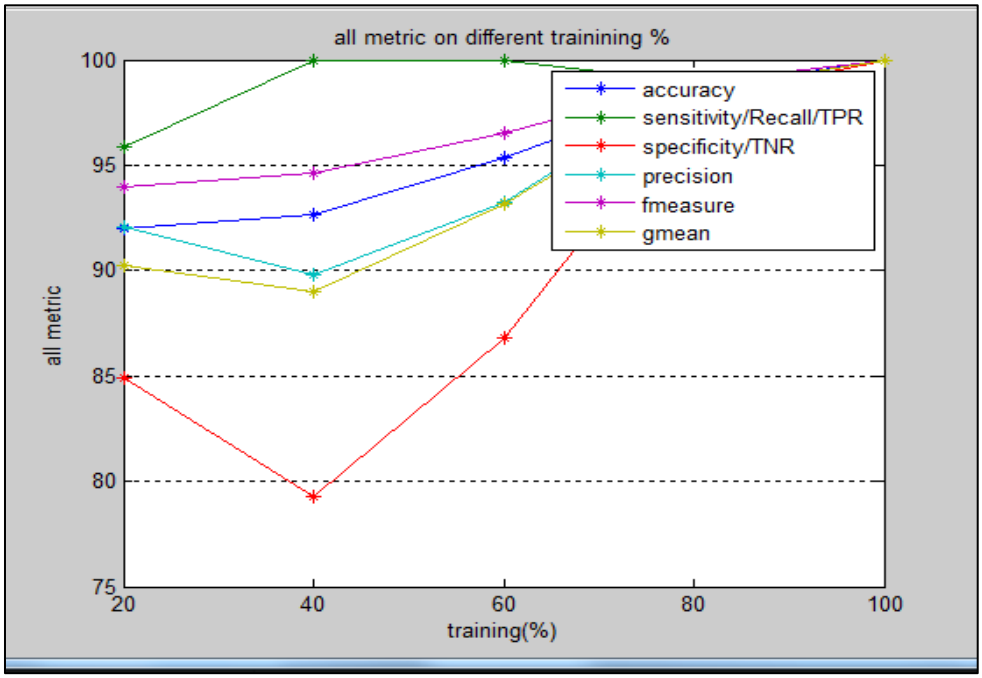

Fig. 1: Graph

V. CONCLUSION AND FUTURE SCOPE

Various techniques have been used to introduce a time efficient and an accurate spam detection system. Results have been obtained on different percentages of the trained data and the respective graph has also been obtained but the features which are identified here are just related to the spam data like no. of URL in the tweet or number of spam words etc. The future researcher can do the work of spammer detection by including the features like user's friend, frequent links etc or can introduce an entire new methodology by using any other classification technique.

REFERENCES

[1] Wikipedia, "Naïve Bayes spam filtering," 2014, http://en.wikipedia.org/wiki/Naive_Bayes_spam_filtering. Accessed on 29-January-2015.
[2] Meheran Sahami, Dumai, S., Heckerman, D., Horvitz, E. "A Bayesian Approach to Filtering Junk E-Mail", AAAI'98 Workshop on Learning for Text Categorization, Madison, Wisconsin. July 1998.
[3] Hassan, T., Cole, P.; Chun Che Fung "An Intelligent Spam Filter-GetEmail5" IEEE Conference on Cybernatics and Intelligent System, Bangkok,7-9 June 2006, pp. 1-5.
[4] Yang Li, Binxing Fang, Li Guo, Shen Wang " Research of a Novel Anti- Spam Technique Based on User's Feedback and Improved Naïve Bayesian Approach", IEEE International Conference on International and Services, Silicon Valley CA, 16-18 July 2006, pp. 86.
[5] Nello Cristianini, John Shawe -Taylor " An Introduction to Support Vector Machines and Other Kernel Based Learning Learning Methods, Cambridge University, 2001, Vol. 22, pp. 103- 104.
[6] Federal Trade Commission, USA: False Claims in Spam, April 30, 2003.
[7] What is spam filter? http://searchmidmarketsecurity.techtarget.com/definition/spam-filter., Accessed on 29-January-2015.
[8] Yu-Fen Chiu, Chia-Mei Chen, BingChiang Jeng, Hsiao-Chung Lin "Alliance based Anti-Spam approach", IEEE Third International Conference on Natural Computation, Haikou, 24-27 August 2007, Volume 4, pp. 203-207.
[9] Cluster, http://searchexchange.techtarget.com/definition/cluster Accessed on 14-August-2015
[10] Data Mining with R , http://www.rdatamining.com/docs/introduction-to-data-mining-with-r Accessed on 14-August-2015
[11] Stop Word, https://en.wikipedia.org/wiki/Stop_words. Accessed on 14-August-2015.
[12] Stemming, https://en.wikipedia.org/wiki/Stemming. Accessed on 14-August-2015
[13] MATLAB, http://in.mathworks.com/products/matlab/ Accessed on 14-August-2015