Experimental Analysis of KNN with Naive Bayes, SVM and Naive Bayes Algorithms for Spam Mail Detection

Divya Sharma
MRIU, Faridabad, Haryana, India

Abstract
Spam is one of the major problem faced in the world of internet. Spam is responsible for flooding the internet with numerous copies of similar messages anonymously which may also contain various attachments that attacks our systems with viruses and are responsible for the origin of botnets and they may also contain phishing emails. In addition time taken by people in reading and deleting spam mails is a waste and a cumbersome task.Spam is a wider term used for junk, fraudulent and unsolicited emails. This research paper consists of comprehensive study of machine learning approaches for spam mail detection such as SVM, KNN and Naive Bayes. These algorithms are among the most influential data mining algorithms in the research community.The detection of fraudulent mails isconsidered as classification problem.In this paper experiments have been performed on different classification methods such as SVM, Naive Bayes and KNN along with Naive Bayes and analysis is done between them.

Keywords
Botnets, Classification, Fraudulent Mails, KNN, Naive Bayes, SVM, Spam

I. Introduction

Emails are one of the easiest mode of communication.It can be classified as Ham and spam. Email is prone to spam mails because it is widely used, less costly and fastest way of communication around the world.A Spam mail is considered as illegitimate as they are not valuable to the user and are unwanted detritus that chokesand clutters their mailboxes. It is also defined as Internet Spam in which majority of messages contains substantially identical content. It incurs high cost for the organizations costs billions of dollars per year to service providers for the loss of bandwidth.According to the Symantec Intelligence Report, 2013 Spam contributes in 71.9% of the email traffic. There can be different intensions of sending spam mails,but major are cyber crimes,phishing and for the purpose of advertisements in the most cost effective manner.Spams can begrouped into two categories. One is cancellable Usenet spam where a single message is sent to twenty or more Usenet groups and suppress the ability of the administrator of that system.Another type of email spam targets individual users with bulk of messages.Machine learning is used to develop automatic spam mail detection system and different machine learning algorithms will be analysed to check accuracy of spam mail detection.

II. Related Literature

Dheeraj Pal, Alok Jain, Aradhana Saxena and Vaibhav Agarwal [1] analysed large amount of data and correlations between them by using Machine Learning Algorithms with the help of WEKA (Waikato environment for knowledge analysis) tool. They have found the prediction value of dataset and data which is stored in different forms like matrix, graph, tree etc by using WEKA ,which consists of different learning techniques for Classification which were implemented in Java.The patterns of the data was analysed by converting it into graphs and visual inspection was done,which the mining software might overlook.They studied the expected trend in the employee's information data of different forms like numeric and nominal, how many of them come under maximum, minimum, mean, and standard deviation.A java program was used to process the data in the software and pull data from the database into data mining format. The converted file can be read like a table which has its own column and helps to organize the data for mining techniques. Later they have shown that it was easy to select the attributes as it is saved in the software, i.e., in the "WEKA" tool. They concluded from the related work that the attribute selection plays an important role to identify parameters that are important and significant for an excellent result. They efficaciously compared the result of all the algorithms with each other Naïve Bayes, J48 algorithm, ROC curve. And thus concluded that the result of J48 tree and the ROC curve is better and easy to understand as compared to Naïve Bayes rule.M. Rathi and V.Pareek [3] have analysed various data mining approach to a spam dataset in order to find out the best classifier for email classification.Initially they have applied classifiers one by one on the entire dataset without selecting features and later they have applied Best-First feature selection algorithm which derives desirable features.campaign and later correlates different data sources such as such as passive DNS, malware, geolocation to provide more insights to spam campaign. Then they had given a score to the spam campaign based on several customizable criteria.By this they had provided a strong platform to conduct investigations on cyber based crime activities. They had applied clustering techniques to generate clusters of emails that are similar or close to each other.Then algorithms such as w-shingling and the jaccard coefficient, Context Triggered Piecewise Hashing (CTPH) or Locality-Sensitive Hashing (LSH) were applied to produce three resemblance scores of all the email pairs in each campaign.Spam campaign detection was done by Frequent –pattern(FP) tree.Spam campaigns were characterised based on different IP address and related host names. Their system had greatly reduced investigation efforts by consolidating spam emails into campaigns.M. Basavaraju and D. Prabhakar [5] have proposed a new spam detection technique that creates clusters of the data based on vector space model. They have divided the data into two groups based on thesimilarity of patterns.The objects are being classified as points or patterns in N-dimensional metric space and similarity between them is measured on the basis of Euclidean distance between pair of points or by calculating cosine of the angle between vectors corresponding to the document. M.Basavaraju have proposed an effective algorithm by consolidating the features of K-means algorithm and BIRCH algorithm. They concluded that K-means clustering algorithm works well for smaller data sets. The combination of BIRCH with K-NNC works better with large data sets. Thus BIRCH is a better clustering algorithm requiring a single scan of the entire data set thus saving time.

Later they applied classification algorithm based on those features. According to their study accuracy is improvedin the results where feature selection is embedded.The authors[4] have elaborated

the methodologies for spam campaign detection,analysis and investigation. They had proposed a framework that integrate spam mails into campaigns, then labels spam campaignsby generating related topics for each

III. Spam Mail Detection

Spam detection is one of the most important interdisciplinary developments in the field of Information technology. It has importance regarding finding patterns, forecasting, discovery of knowledge etc. The manual analysis of spam data is impractical and dubious due to it's astronomic size. There are various legal measures that are adapted for detecting spam mails but they have limited effect. Automatic Email spam classification also contains challenges because of Unstructured information, large size of documents and more number of features. Detection of Spam mail is a classification task and promising classification can be achieved by the selection of effective algorithm. Anti-spam filters, software tools such as Spambayes which is used by Microsoft outlook, SpamAssassin System, SpamBouncer or Mozilla Junk Mail Control [14] have more direct value and attempt to block spam messages automatically but they still exhibit some problems as these filters rely on manually constructed Keyword patters. To be most effective and to avoid deletion of non-spam messages the Keyword patters needs to be manually tuned. But it is time effective and requires expertise that is not always available. The worst part is the characteristics of spam messages change over time for which Keyword patterns needs to be updated frequently. Some of the effective techniques for mail classification are discussed below.

IV. Support Vector Machine

Support Vector Machines, SVM[7] is considered as state-of-art classification method for text categorization.It is a predictive model which takes the input data and generates the output and thus classifies the data into two categories. SVMtraining algorithm can be best implemented by building a model for those text corpus where each training example belongs to one of the two classes [3]. Then the data is divided into two categories by the construction of N-Dimensional hyperplane. Two parallel hyper planes are constructed on each side of hyper plane that separates the data where the separating hyper plane maximises the distance between two hyper planes. A linear classification function is generated for a linearly separable dataset corresponding to a separating hyperplane f(X) that passes through the middle of the two classes and separates them[6]. After determining this function, a new data instance Xn can easily be classified by testing the sign of the function f(Xn);

Where Xn belongs to a positive class if f(Xn) > 0

The generalization error of the classifier will be better for larger distance or margin.It can work well on high dimensional feature set and can transform non-linearly separable data to a new linearly separable data by using kernel trick [10]. SVM can be easilyextended to perform numerical calculations and it can be used to conduct Regression analysis[6] .It can also be used to rank the elements and it is insensitive to the outliers but choice of the Kernal can be the tedious task.

V. KNN Classifier

K-nearest neighbour is a sophisticated approach for classifica--tion that finds a group of K objects in the training documents that are close to the test value. To classify an unlabeled object, the distance between this object and labelled object is computed and it's K nearest neighbours are identified.Classification accuracy mainly depends on the chosen value of K and will be better than that of using the nearest neighbour classifier[5]. For large data sets, K can be larger to reduce the error. Choosing K can be done experimentally, where a number of patterns taken out from the training set can be classified using the remaining training patterns for different values of k.The value of K which gives the least error in classification will be choosen. If same class is shared between several of K-nearest neighbours, then per-neighbour weights of that class are added together, and the resulting weighted sum is used as the likelihood score of that class with respect to the test document [8]. A ranked list is obtained for the test document by sorting the scores of candidate classes.Decision rule i.e Score(d,ci) for KNN can be written as:

$$\sum_{d_j \epsilon KNN(d)} \mathrm{Sim}(\mathrm{d}, \mathrm{d_j})\delta(\mathrm{d_j}, \mathrm{c_i})$$

Where d is the test document, C_i indicates the classes of KNN which is used by the system to find K-nearest neighbours among training documents, KNN(d) is the set of K-nearest neighbours of document d, $d(d_j, c_i)$ is the classification for do-cument d_j with respect to class c_i, that is the value of $\delta(d_j,c_i)$ Will be 1 if dj is an element of class c_i, Else it will be 0 For the test document d, it should be assigned the class that has highest resulting weighted sum. The classification of KNN is easy to understand and implement and it can perform well in many situations. It is also scalable to new modifications as it is possible to eliminate manyof the stored data objects, but still retain the classification accuracy of the KNN classifier. This is known as 'condensing' and can greatly speed up the classification of new objects but there comes the difficulty while deciding the value of K. If K is too small then result can be sensitive to noise points whereas if for large value of K, the neighborhood may include too many points from other classes. The choice of the distance measure is another important consideration [6]. Although various measures can be used to compute the distance between two points, but smaller distance between two objects does not always implies a greater likelihood of having the same class.

VI. Naïve Bayesian Classifier

Sahami [13] discussed a Machine Learning algorithm to build a filter which processes previously received spam and legitimate messages and on the basis of that it learns how to block incoming spam messages automatically, to deal with the problem of continuously changing Keyword patterns which needs to be updated periodically. In the context of text classification it is necessary to represent mail messages as feature vectors to make Bayesian Classification methods directly applicable [12]. The Naïve Bayesian classifier assumes that each document is represented by a vector x.Let x be the vector of values from x_1 to x_n. Where $x_1...x_n$ are the values of the attributes $X_1....X_n$ in the vector space model. Following Sahami et al. binary attributes are used i.e X1 = 1 if message has the property represented by Xi otherwise X1 = 0.Along with that the property of Mutual Information(MI) is used to select among all possible attributes. MI(X;C) is calculated as:

$$\sum_{x\epsilon\{0,1\},c\epsilon\{spam,legitimate\}} P(X = x, C = c) \cdot \log \frac{\mathrm{P}(\mathrm{X} = \mathrm{x}, \mathrm{C} = \mathrm{c})}{\mathrm{P}(\mathrm{X} = \mathrm{x}) \cdot \mathrm{P}(\mathrm{C} = \mathrm{c})}$$

Where X is the attribute with category denoting variable C. Attributes of highest Mutual Information, MI values is selected which is time consuming. The probability of P(X|C), P(C) and P(x)

is estimated by Frequency ratios. Baye's Theorem and theorem of total probability is used to determine the probability, that a document with the vector of x equals to ‹x_1,x_2,x_3,….x_n› belongs to a category c, which is given below:

$$P\left(C=c|\overrightarrow{X}=\overrightarrow{x}\right)=\frac{\mathrm{P}\left(\mathrm{C}=\mathrm{c}\right)\cdot\mathrm{P}\left(\overrightarrow{X}=\overrightarrow{x}\mid\mathrm{C}=\mathrm{c}\right)}{\sum_{k\epsilon\{spam,legitimate\}}\mathrm{P}\left(C=k\right)\cdot\mathrm{P}\left(\overrightarrow{X}=\overrightarrow{x}|C=k\right)}$$

Here probability that a class C belongs to X attribute i.e P(X|C), is impossible to find because there are too many possible values of X.So, Naive Bayesian classifier allows us to compute the probability that attribute, X belongs to class C as below:

$$\frac{\mathrm{P}\left(\mathrm{C}=\mathrm{c}\right)\cdot\prod_{i=1}^{n}\mathrm{P}\left(X_i=x_i|C=c\right)}{\sum_{k\epsilon\{spam,legitimate\}}\mathrm{P}\left(\mathrm{C}=\mathrm{k}\right)\cdot\prod_{i=1}^{n}\mathrm{P}\left(\mathrm{X_i}=\mathrm{x_i}|\mathrm{C}=\mathrm{k}\right)}$$

where it is assumed that X_1…X_n are conditionally independent with category C.So,it becomes possible to compute P(C|X). The Naïve Bayes model is tremendously appealing due to it's simplicity and androbustness [11].

VII. Experimental Analysis

In order to validate the proposed schemefor spam mail detection several experiments are conducted. The main objective is to find out the best classifier whose accuracy is better than the rest of the classifiers. Spam base dataset TREC 2007 public corpus is used. It consists of 12 attributes and 4899 messages.

The classification algorithms that are applied one by one on the dataset are:Naïve Bayes,Support Vector Machine and KNN with Naïve Bayes. And then Fmeasure is calculated with respect to recall and Accuracy value for different percentage of training dataset taken.After comparing all the three classifiers it is concluded that the Fmeasure for modified classifier i.e KNN with Naïve Bayes is maximum followed by SVM andand it is minimum for Naïve Bayes classifier for different recallvalues.Similarly, the accuracy measure also follows the same sequence when calculated for different percentage of training data set taken .ie it is maximum for KNN with Naïve bayes classifier followed by SVM and minimum for Naïve Bayes classifier.

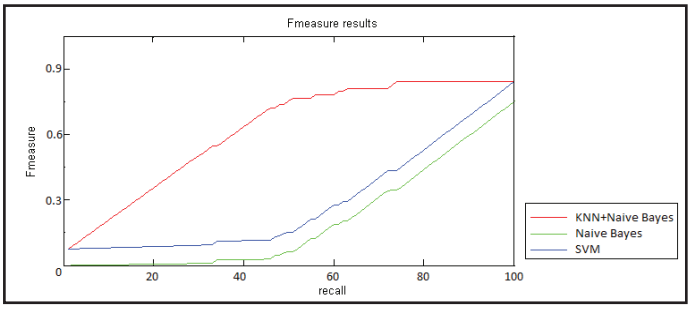

Fig. 1: Plot of Fmeasure with Respect to Recall

Table 1: Fmeasure with Respect to Recall Value for Different Classifiers

Recall (in %)	Fmeasure for Naïve Bayes	Fmeasure for SVM	Fmeasure for KNN+Naïve Bayes
20	0.001	0.100	0.325
40	0.023	0.127	0.687
60	0.157	0.248	0.814
80	0.482	0.536	0.848
100	0.689	0.848	0.848

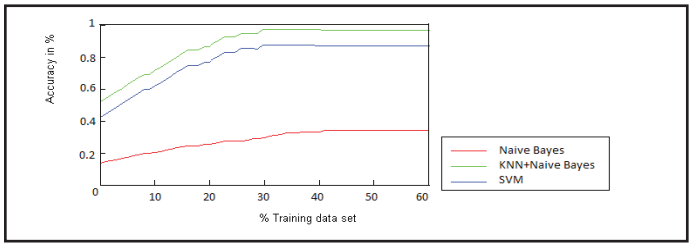

Fig. 2: Accuracy Measure for Different % of Data Sets Taken

Table 2: Accuracy Measure for Different % of Training Dataset Taken for Different Classifiers

Training Data Set (in %)	Accuracy for Naïve Bayes	Accuracy for SVM	Accuracy for KNN+Naïve Bayes
10	0.200	0.591	0.699
20	0.257	0.774	0.875
30	0.324	0.947	0.837
40	0.356	0.887	0.989
50	0.378	0.887	0.989
60	0.378	0.887	0.989

VIII. Conclusion and Future Scope

Researches are done to find out the best classifier for spam mail detection. So various classification algorithms are applied on the given input data set and the results are checked. In this research paper KNN classifier is combined with Naïve Bayes classifier. Fmeasure is calculated with respect to recall and Accuracy is found with respect to different percentage of training data sets taken.The modified classifier is compared with Naïve Bayes and SVM algorithm. KNN is combined with Naïve Bayes to make the classification more accurate.Where K nearest neighbours are found first and then Naïve Bayes algorithm is applied which is already considered as one of the oldest and efficient classifier for detecting spam mails. In future various other Genetic algorithms can be combined to make the pre-existing classifiers more efficient.

References

[1] Dheeraj Pal, Alok Jain, Aradhana Saxena, Vaibhav Agarwal, "Comparing Various Classifier Techniques for Efficient Mining of Data", Proceedings of the International Congress on Information and Communication Technology, pp. 191-202, 2016.

[2] B. U. Gaikwad, P. Halkarnikar, "Spam E-Mail Detection by Random Forest Algorithm", International Journal of Advanced Computer Engineering and Communication Technology (IJACECT), Vol. 2, No. 4, 2013, pp. ISSN (Print):2319-2526, 2013.

[3] M. Rathi, V. Pareek,"Spam Mail Detection through Data Mining – A Comparative Performance Analysis", IJMECS, Vol. 5, No. 12, pp. 31-39, 2013.

[4] S. Dinh, T. Azeb, F. Fortin, D. Mouheb, M. Debbabi, "Spam campaign detection, analysis, and investigation", Digital Investigation, Vol. 12, pp. S12-S21, 2015.

[5] M. Basavaraju, D. Prabhakar,"A Novel Method of Spam Mail Detection using Text Based Clustering Approach", International Journal of Computer Applications, Vol. 5, No. 4, pp. 15-25, 2010.

[6] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, Z. Zhou, M. Steinbach, D. Hand, D. Steinberg,"Top 10 algorithms in data

mining", Knowledge and Information Systems, Vol. 14, No. 1, pp. 1-37, 2007.
[7] S. Nizamani, N. Memon, M. Glasdam, D. Nguyen, "Detection of fraudulent emails by employing advanced feature abundance", Egyptian Informatics Journal, Vol. 15, No. 3, pp. 169-174, 2014.
[8] S. TAN,"Neighbor-weighted K-nearest neighbor for unbalanced text corpus", Expert Systems with Applications, Vol. 28, No. 4, pp. 667-671, 2005.
[9] P. Sao, P. Prashanthi,"Email Spam Classification Using Naive Bayesian Classifier", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Vol. 4, No. 6, 2015.
[10] Joachims T. ,"A statistical learning learning model of text classification for support vector machines", Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, ACM, pp. 128–36, 2001.
[11] T. Guzella, W. Caminhas,"A review of machine learning approaches to Spam filtering", Expert Systems with Applications, Vol. 36, No. 7, pp. 10206-10222, 2009.
[12] S. Hooda, V. Kansal, S. Kadian,"Comparison and Analysis of Spam Detection Algorithms", International Journal for Research in Applied Science & Engineering Technology (IJRASET), Vol. 3, No. 1, 2015.
[13] M. Sahami, S. Dumais, D. Heckerman, E. Horvitz,"A Bayesian Approach to Filtering Junk -Email.In Learning for Text Categorization", AAAI Workshop, pp. 55-62, 1998.
[14] P. Chirita, J. Diederich, W. Nejdl,"MailRank: Using Ranking for Spam Detection", CIKM'05, ACM, 2005.
[15] T. Wu, K.-T. Cheng, Q. Zhu, Yi-L. Wu,"Using visual features for anti-spam filtering", In Proceedings of the IEEE International Conference on Image Processing, Vol. III, pp. 501–504,2005.
[16] E. Blanzieri, A. Bryl,"A survey of learning-based techniques of email spam filtering", Artificial Intelligence Review, Vol. 29, No. 1, pp. 63-92, 2008.
[17] D. DeBarr, H. Wechsler,"Spam Detection using Clustering, Random Forests and Active Learning", CEAS 2009 – Sixth Conference on Email and Anti-Spam, 2009.
[18] M. Hasnat, O. Alata, A. Trémeau,"Model-based hierarchical clustering with Bregman divergences and Fishers mixture model: application to depth image analysis", Statistics and Computing, Vol. 26, No. 4, pp. 861-880, 2015.
[19] J. Chen,"Making clustering in delay-vector space meaningful", Knowledge and Information Systems, Vol. 11, No. 3, pp. 369-385, 2006.
[20] S. Cost, S. Salzberg,"A weighted nearest neighbor algorithm for learning with symbolic features", Mach Learn, Vol. 10, No. 1, pp. 57-78, 1993.
[21] K. Li, Y. Zhang, Z. Li,"Application Research of Kalman Filter and SVM Applied to Condition Monitoring and Fault Diagnosis", AMM, Vol. 121-126, pp. 268-272, 2011.