Adaptive Approach for Spam Detection

Sumant Sharma[1], Amit Arora[2]

[1] Student, Department of Computer Science, Chitkara University, Himachal Pradesh.

[2] Assistant Professor, Department of Computer Science, Chitkara University, Himachal Pradesh.

Abstract

Spam has emerged as a major problem in recent years. The most widely recognized form of spam, is email spam. The accounts which contain spam messages must waste time deleting annoying and possibly offensive message. In this paper, we present a variety of machine learning algorithms to identify spam in e-mail accounts. We design classifier model to automatically determine spam in the accounts so that time of account holder can be saved and utilized on other work.

The dataset we used for our project is named as SPAMBASE dataset download from UCI Machine Learning Repository. We used the labeling data in conjunction with machine learning techniques provided by WEKA tool kit, to train a computer to recognize spam instances automatically. The accuracy of 94.28 is shown by the Random committee through the experiment.

Keywords: *Machine Learning, NumericToBinary Filter, Spam Detection, Weka.*

1. Introduction

Spam is the use of electronic messaging systems to send unsolicited bulk messages, especially advertising, indiscriminately [1].

While the most widely recognized form of spam is e-mail spam. Now a days spam messages are sent to large number of users via internet. The Spammers are targeting the users having accounts on e-mail sites like Gmail, Hotmail, I Cloud and social networking sites like Facebook and Twitter. An account holder waste time deleting annoying and possibly offensive message, It also cause delay to deliver important e-mails to the account due to large amount of spam traffic in between host and the e-mail servers. Hence, filtering the spams from bulk of data is very challenging task. One of the popular method for spam detection is Bayesian spam filtering (Thomas Bayes) which is a statistical technique for e-mail filtering. In this technique a naïve Bayes classifier is used to identify spam e-mail. Bayesian classifier work by correlating the use of tokens with spam and non-spam e-mails and then using Bayesian inference to calculate a probability that an e mail is spam or not [8].

Another approach is content based spam filtering in which a word or a phrase of an email is analyzed by using machine learning algorithm. It matches the content, if the content is found then numeric value is given to the e-mail. A threshold value is set and after crossing a threshold, that e-mail is considered as spam [4].

Another approach uses Support Vector Machine algorithm for content base filtering. This algorithm gives a remarkable performance on text classification also. Support Vector Machine algorithm give very high performance when applied on large benchmark dataset. An equivalent performance was also evaluated by applying the Relaxed Online SVM (ROSVM) on same dataset to detect E-mail spam, Blog spam and SP-LOG [6].

Another approach detects spam by using the text clustering based on vector space model. The disjoint clusters are computed for all Spam or Non-Spam mails by using the spherical k-means algorithm. For each centroid vectors, label is assigned by calculating the number of spams in the cluster. When new mail arrives in the account, the cosine is calculated between the new mail vector and centroid vector. Finally, the label of the most relevant cluster is assigned to the new mail [7].

In this paper SPAMBASE dataset is used to classify e-mail as spam or non-spam e-mails. Spam base dataset is multivariate dataset contains data from a single email account. This data is used to apply various machine learning algorithm to classify the spams present in that. For the various machine learning algorithm **WEKA** tool is used. WEKA is open source software made in java. It provides collection of algorithms used for data analysis and predictive modeling. After applying the algorithm, percentage of precision, recall accuracy, score and Correctly Classified Instances at ten Fold Cross-validation is calculated. The classifier which is having high accuracy and correctly classified instances.

2. Experiment

In this section, we describe the experiment done in order to detect spam. How we apply the machine learning algorithm and which algorithm is giving the maximum number of correctly classified instances and accuracy. We will draw our attention on detecting maximum number of spams from the spam-base dataset. In particular we look for the following-

[1] Bayes Net (BN)
[2] Logic Boost (LB)
[3] Random Tree (RT)
[4] JRip (JR)
[5] J48 (J48)
[6] Multilayer Perceptron (MP)
[7] Kstar (KS)
[8] Random Forest (RF)
[9] Random Committee (RC)

Bayes network model is a probabilistic graphical model that represents a set of random variable and their conditional dependencies via a directed graph. For example a Bayesian network could represent the probabilistic relationships between diseases and symptoms [9].

Random forests are an ensemble learning method for classification (and regression) that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. The algorithm for inducing a random forest was developed by Leo Breiman and Adele Cutler [10].

J48, In machine learning are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification analysis. The basic j48 takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier [11].

Multilayer Perceptron (MLP) is a feedforward artificial neural network model that maps sets of input data onto a set of appropriate outputs. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron with a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training the network [12].

To do all the experiments, we used WEKA (Waikato Environment for Knowledge Analysis), the open source software to make and test the classifier. WEKA is a popular suit of machine learning algorithms for data mining tasks. It is widely used for developing new machine learning models. WEKA has a highly customizable interface and very easy to use, which enabled us to run a large number of experiments.

For all the experiments, we used a Spam-base dataset. This dataset was collected from UCI Machine Learning Repository for the research purpose. Dataset is multivariate having 4601 instances and 55 attributes, consisting of tagged emails from a single email account. We have a training set of labeled spam mails to train the classifiers. Testing of the classifier was performed on a testing set, the performance was measured by evaluating the accuracy, precision, recall and score for various classifier.

2.1 Preprocessing Data

The data which is available in the spam-base data set is in numeric form. The fifty five attributes in the dataset represent relative frequencies of various salient words and characters in emails. We wish to convert these to Boolean values for the experiment. The attribute will take a value 1 if the word or character is present in the email and 0 if it is not present in the email. To do this we apply a Numeric to Binary filter present in the WEKA tool. This filter will convert all the numeric values to the binary

2.2 Making Classifier

The converted data set is used to train the classifier to detect spam from regular email by checking the number of occurrences of each word for all the spam and non-spam e-mails. A variety of algorithm is given into the WEKA tool that can be used. We apply the algorithm by choosing “TEN Fold Cross Validation”.

3. Examing Results

Now, we examin results of classifier/model prodeced by the weka. Which basically tells Five things which are correctly classified instances(CCI), True positive value(TP), False positive value(FP), True negative value(TN), False negative value(FN). For each classifier all these five thing are distant. All this five values were used to calculate Accuracy, Recall Precision and Score for a particular classifire.

TABLE-1

Algorithm	CCI	TP	FP	TN	FN
NB	88.54	2596	192	1478	335
BN	88.56	2596	192	1479	334
NBU	88.54	2596	192	1478	335
LOGISTIC	92.95	2654	134	1623	190
MLP	93.28	2630	158	1662	151
SGD	93.28	2655	133	1637	176
SMO	93.21	2659	129	1630	183
VP	92.56	2615	173	1644	169
KSTAR	93.56	2665``	123	1640	173
DT	91.71	2666	122	1554	259
RIP	92.32	2634	1544	1614	199
PATR	93.06	2631	157	1651	162
DS	77.2	2041	747	1511	302
J48	92.34	2618	170	1631	182
RF	93.89	2673	115	1647	166
RT	91.54	2586	202	1626	187
BAGGING	92..93	2650	138	1626	187
LOGICBOOST	89.76	2589	199	1541	272
MCC	92.95	2654	134	1623	190
RS	92.37	2667	121	1583	230
CVR	92.15	2637	151	1603	210
FC	92.34	2618	170	1631	182
RC	94.28	2680	108	1658	155

As we can see in the above table each individual classifier is having five values. These values are used for further calculattions.

4. Indentation And Equations

Now, Precision, Recall, Accuracy and Score are calculated for each of the individual classifier. To calculate we need to have the values from Figure 1. From the figure rather than calculating values for all the classifiers we took some of the classifiers based on the percentage of correctly classified instances (CCI). After that the future calculation is done by using the formulas describe below-

Precision (PPV) =TP/ (TP + FP)

Recall (TPR) = TP/ (TP + FN)

Accuracy (ACC) = (TP + TN) / (TP + FN)+(FP +TN)

Score(F1) = 2*TP/ (2*TP + FP + FN)

TABLE-2

ALGOS	CCI	PPV	TPR	ACC	F1
BAYSNET	88.56	93.113	88.6	88.56	90.8
LOGICBOOST	89.76	92.862	90.4	89.7	91.66
RANDOMTREE	91.71	92.754	93.2	91.54	93
JRIP	92.32	94.476	92.9	92.32	93.71
J48	92.34	93.902	93.5	92.34	93.7
MULTILAYER PERSAPTRON	93.28	94.332	94.5	93.28	94.45
KSTAR	93.56	95.588	93.9	93.56	94.73
RAMDOM FOREST	93.89	95.875	94.1	93.89	95
RANDOM COMMITTEE	94.37	96.126	94.5	94.28	95.32

Table- 2 describes all the algorithms with their calculated values of percision, recall, accuracy and score. All the values are calculated manually by using the formulas describe above. Here percentage values is taken for all the algorithms. As we can see the accuracy of the algorithm is lies between 88.56 to 94.5.

Figure-1

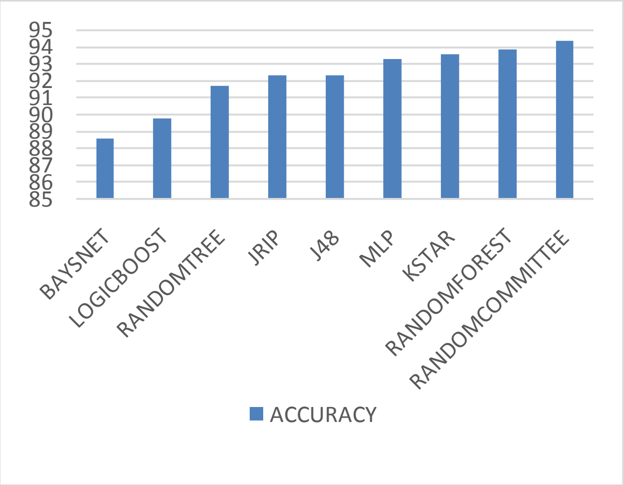

Figure-1 shows Accuracy of 9 algorithm on spambase dataset.

Figure-2

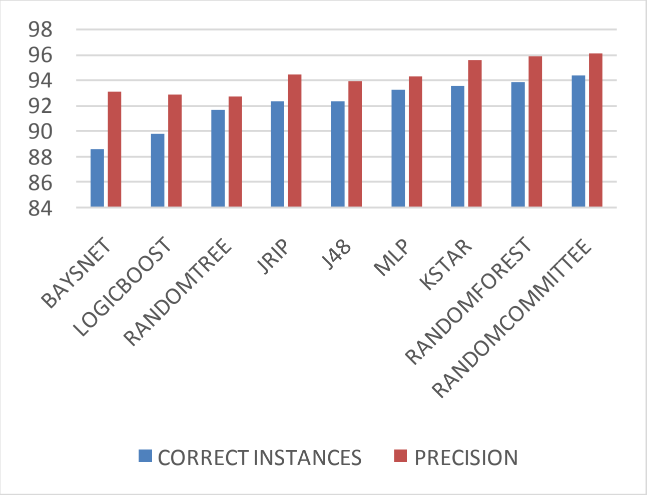

Figure-2 shows Correctly Classified Instances and Percision after appling the classifier on spam base dataset in percentage.

Figure- 3

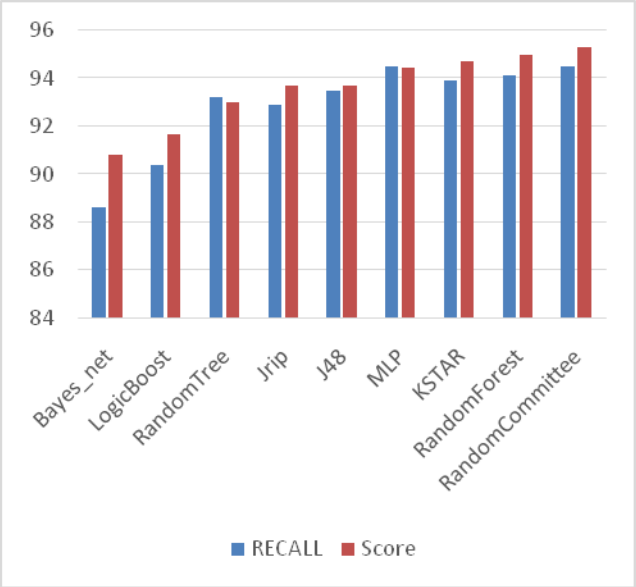

Figure- 3 shows the Recall and Score value in percentage.

5. Conclusions

In this paper we apply various machine learning algorithms to classify the spams from e-mail. We compare 24 algorithms (Table 1) on spam base dataset using 55 spam base-attributes. We focus on the accuracy and performance of the algorithm to classify the spam/non-spam e-mails from tagged emails of a single account. As an initial experiment to analyze accuracy and performance 10 classifiers are used as described in the Table-2 above. The results of experiencing on Spambase datasets show better performance of the proposed method. It shows 94.28 accuracy and Score value of 95.32 for Random Committee, which is high in all other methods

Acknowledgments

I would like to express a deep sense of gratitude and thanks profusely to Mr. Amit Arora, Assistant professor, Department of Computer Science and Engineering, COSET who was the thesis Supervisor. Without the wise counsel and able guidance, it would have been impossible to complete the research work in this manner and on time.

References

[1] Guzella T. S., Caminhas W. M. 2009"A review of machine learning approaches tospamfiltering". 36(7), 10206-10222.

[2] Koprinska I., Poon J., Clarck J., Chan J. "Learning to classify e-mail". Info. S. 177: 2167-2187, 2007.

[3] Kuncheva L. "Combining Pattern Classifiers, Methods and Algorithms, Wiley Inter Science, 2005".

[4] S. Mason. New Law Designed to Limit Amount of Spam in e-mail.

[5] Jongsub Moon, Taeshik Shon, Jungtaek Seo, Jongho Kim, Jungwoo."An Approach for Spam E-mail Detection with Support Vector Machine".

[6] D.Sculley and Gabrial m.wachman. "Relaxed Online SVMs for Spam Filtering".

[7] Spam detection using text clusteringSasaki, M.; Dept. of Computer& Inf. Sci., Ibaraki University; Shinnou, H.

[8] M. Sahami, S. Dumais, D. Heckerman, E. Horvitz (1998). "A Bayesian approach to filtering junk e-mail ".

[9]Pearl,Judea(2000).Causality:Modelsasoning,and Inference. Cambridge University Press. ISBN 0-521-77362-8.

[10]Breiman,Leo (2001),"RandomForests". Machine Learning 45 (1): 5–32.10.1023/a: 1010933404324

[11] Cortes, Corinna; and Vapnik, Vladimir N.; "Support-Vector Networks", Machine Learning, 20, 1995.

[12] Rumelhart, David E., Geoffrey E. Hinton "Learning Internal Representations & Error Propagation".