Comparative Study On Supervised Machine Learning Algorithms For Spam Mail Detection

C.Nalini, R.Shantha Kumari,J.Sudeeptha

Abstract: Electronic mail (E-mail) is used to exchange messages between people via internet. E-mail protocols like Simple Mail Transfer Protocol (SMTP), POP (Post Office Protocol) and IMAP (Internet Message Access Protocol) are used to transfer messages from sender to receiver. Due to the flaws in E-mail protocols, development of online businesses and advertisement companies create E-mail based intimidation. E-mail spam is called as junk mail. Today handling spam mail is one of the major problems in software companies. Since spam mail causes traffic problems and bottle necks that limit memory space, computing power and speed. And also a user has to spend more time to detect and obliterate spam mails. Machine learning models are used to are used to overcome this problem. Machine learning models are categorized into supervised, unsupervised and semi supervised learning models. Supervised learning models are used to classify E-mails, filter and prevent the spam mail. The proposed work explores the performance of machine learning algorithms like Decision Tree(DT), Navie-bayes, k-Nearest Neighbours (k-NN),Support Vector Machines(SVM) and Random Forest(RF) learning algorithms for classifying spam messages from E-mail. Accuracy, F-measure and recall parameters are used to evaluate the performance of the learning algorithms.

Key words: Spam mail,Random Forest, Navie-bayes,k-NN,SVM,DT, Ensemble learning algorithm

1. INTRODUCTION

E-mail dominates business world for sending bulk of messages to customers at low cost. But,the cybercriminals use mail as the weapon to reduce the productivity of a competitive companies. Knowledge engineering approach (i.e) rule based knowledge system use to detect the incoming mail is legitimate or not. But database require continuous updating. Machine learning algorithms are used to build an efficient decision model for detecting spam E-mail. But build a machine learning model is easy. So, today researchers concentrate on machine leaning algorithm to develop an efficient spam mail detection system. Spam filtering methods are used to filter spam E-mail. However, there exists a risk of misclassification or removal of legitimate email as spam.

2.RELATED WORK

Many researchers and academicians have proposed different types of classification algorithms to build a decision model. Classification models are used to classify data into groups based on class labels. The volume of spam E-mails are increased day by day. We need efficient automatic spam E-mail detection. Traditional data analysis techniques are not suitable to construct an efficient detecting system. The objectives of this work is to compare different types of classification algorithms and identifies an appropriate algorithm to build a classification system. Puniskis et al. [1] proposed neural network based classification algorithm for spam mail detection.. The result demonstrates that the algorithm perform well ,but ANN based algorithm can't work alone as a spam filter.Patil et al. [2] proposed Navie Bayes classifier to classify E-mail and use bag of words for feature extraction.Li et al. [3] proposed a spam filtering model using ensemble learning algorithms.Sanz et al. [4] presented the research issues related to email spams. The research work illustrated the basic concepts of machine learning approaches suitable filtering email spams.

- *Department of Information Technology,Kongu Engineering College,Erode*
- *[2]Software Engineer,Accenture,Chennai*

2.1 Decision tree algorithm

Ross Quinlan developed a decision tree algorithm known as ID3.C4.5 is a successor of ID3 algorithm. It is a benchmark algorithm for a newer supervised learning algorithm. It is based on greedy approach in which decision tree is constructed in a top-down recursive divide and conquer manner. Information gain attribute selection measure is used to select the splitting attribute.

Decision tree Algorithm for spam E-mail detection:
Input: E-mail data set(D)
Output: classify the E-mail messages into spam/non spam

1. Create a node N, represent tuples in D
2. While (attribute list is empty)
3. For each attribute Calculate information gain for each attribute
4. Select the attribute with the largest gain value as the root.
5. End while
6. Return decision tree model

Attribute selection method:
Entropy and gain search formula as follows:

$$Gain(s) - Entrophy(s) - \sum_{i=1}^{n} \frac{|s_i|}{|s|} \qquad (2.1)$$

Where,
S = the set of cases
n = number of partitions
$|S_i|$ = number of cases in the partition i
|S| = number of cases in S.

$$Entrophy(s) = \sum_{i=1}^{n} - p_i \log_2 p_i \qquad (2.2)$$ where

,
S = The set of cases
n = number of partitions
p_i = proportion of S_i to S.

2.2 Navie Bayes classifier

The Naive Bayesian classification algorithm is a probabilistic classifier. It is based on Bayes' theorem. It applies Bayes' theorem to predict the class label for the unknown record. It is very simple and easy to implement. It

can be used for both binary and multiiclass classification problems.
Bayes' theorem

$$P(spam \mid word) = \frac{P(spam).P(word|spam)}{P(spam).P(word|spam) + P(non-spam).P(word|non-spam)} \quad 2.2.$$

Where,
P(spamword) is probability of a spam word occurred in the E-mail
P(spam) is probability of given message is spam.
P(wordspam) is probability of spam word appears in spam message.
P(non- spam) is the probability of non-spam word
P(wordnon-spam) is the probability of a word occur in non-spam message.

The Navie Bayesian classifier predicts that tuple X belongs to the class C_i if and only if $P(C_i \mid X) > P(C_j \mid X)$ for 1<=j<=m, j≠i

2.3 k-NN algorithm:
k-NN can be used for both classification and regression predictive problems. In k- nearest neighbors, the training data are stored in an n-dimensional pattern space. To predict the class label to the given unknown data,it search pattern space to find k nearest neighbors to the given unknown data and assign the most common class among its k-nearest neighbors. Eculidean distance is used to find the similarity.Min-max normalization is applied before find the similarity between data points.

2.4 Support Vector Machine(SVM):
Support vector machine is used to classify linear and nonlinear data.It transforms original training data into higher dimension using nonlinear mapping and find a hyperplane for data separation.The SVM finds hyperplane using support vectors and margins which is defined by support vectors.

2.5 Random forest:
Random forest is an ensemble learning method.It is a collection of decision tree classifiers. Attributes are randomly selected to determine the split. Each tree model depends on the random selection. Random forest algorithm is more robust to errors and outliers. The accuracy of a random forest depends on the strength of the individual classifiers and dependence between them.

3. RESULT ANALYSIS

To analysis the performance of the algorithms, Spam dataset is downloaded from UCI Machine Learning Repository. The data consists of 4601 instances and each instance has 58 attributes. The attributes represent the frequency of a given word ,character in an instance. In 58 attributes, 48 attributes describes the frequency of word w(i.e) the percentage of words in the email, 6 attributes depicts the frequency of a character c, and 3 attributes illustrates the longest length, total numbers of capital letters and average length. It has two class label such as spam or no spam. WEKA is a Java based open source tool, is a collection of machine learning algorithms. The performance comparative analysis among machine learning algorithms for classifying E-mail messages are done through WEKA tool. Table 3.1,Fig.1,Fig. 2 illustartes the performance of machine learning algorithms.

Metrics for evaluating the classifiers performance
True Positive(TP): The number positive tuples are correctly labeled as a positive tuple.
True Negative(TN): The number negative tuples are correctly labeled as a negative tuple.
False Positive(FP): The number negative tuples are incorrectly labeled as a positive tuple.
False Negative(FN): The number positive tuples are incorrectly labeled as a positive tuple.
Accuracy of a given classifier on a given data set is the percentage of test data tuples are correctly classified.

$$Accuracy = \frac{TP + TN}{P + N} \quad (3.1)$$

Precision: used to measure exactness

$$\Pr ecision = \frac{TP}{TP + FP} \quad (3.2)$$

Recall: used to measure completeness

$$\text{Re}\, call = \frac{TP}{P} \quad (3.3.)$$

F-measure give equal weight to precision and recall

$$F - measure = \frac{2 * precision * recall}{precision + recall} \quad (3.4)$$

Receiver Operating Curve(ROC) is visual tool for comparing two classification models.It shows the trade-off between the True Positive Rate(TPR) and False Positive Rate(FPR)

Table 3.1 *Performance analysis of accuracy*

Algorithm	Accuracy	Precision	F-measure	Recall
Navie Bayes	79.29	0.842	0.793	0.794
SVM	90.42	0.905	0.904	0.903
k-NN	90.78	0.908	0.908	0.908
C4.5	92.98	0.930	0.930	0.930
Random forest	95.50	0.955	0.955	0.955

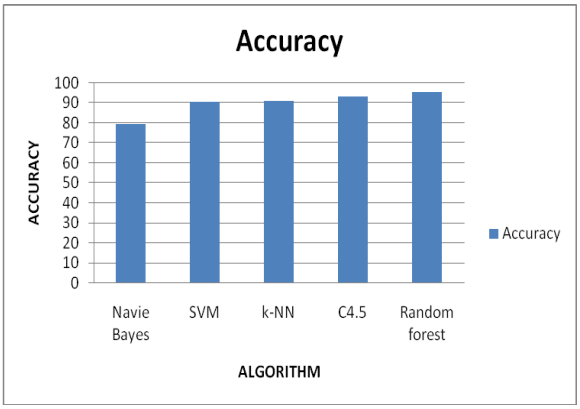

Fig 1. *Analysis accuracy*

Table 2*.Performance analysis of error rate*

Algorithm	ROC Area	RMSE	TP Rate	FP Rate
Navie Bayes	0.937	0.4527	0.793	0.152

SVM	0.891	0.3096	0.904	0.122
k-NN	0.906	0.3036	0.908	0.103
C4.5	0.939	0.2562	0.930	0.078
Random forest	0.987	0.1947	0.955	0.054

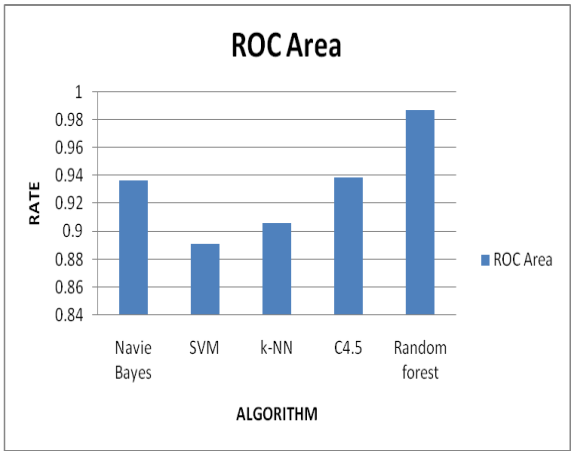

Fig 2. *ROC rate analysis*

The results illustrated that random forest algorithm detect spam mail better than Navie bayes,SVM, k-NN,C4.Simple Navie bayes algorithm misclassified more instances than other algorithms. Although SVM and k-NN algorithm classified 90% of instances classified correctly, root mean square rate of SVM is higher than k-NN. The results convey that the ensemble classification algorithm (Random forest) produce better predictive model than individual classification algorithm.

4. CONCLUSION

Today industries require a good spam mail detection method to filter spam mails. The results demonstrated that ensemble classifier model perform well than a single classifier model. In future heuristics are applied to enhance the performance of spam classification system.

.

REFERENCES

[1] D. Puniškis, R.Laurutis and R. Dirmeikis “An Artificial Neural Nets for Spam

[2] e-mail Recognition”, Electronics and electrical engineering, Vol. 69, No. 5, pp. 73 – 76, 2006.

[3] Patil, T. and Sherekar, S. ”Performance Analysis of Navie Bayes and Classification Algorithm for Data Classification", International Journal Of Computer Science And Applications, 2013.

[4] W. Li, N. Zhong, Y. Yao, J. Liu, C. Liu, “Spam filtering and email-mediated applications”, International Workshop on Web Intelligence Meets Brain Informatics, 2006.

[5] A. Bhowmick, S.M. Hazarika, “Machine Learning for E-Mail Spam Filtering: Review, Techniques and Trends”, arXiv:1606.01042v1 [cs.LG] 3 Jun 2016, 2016, pp.1–27.