International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-1, Issue-6, January 2013

Artificial Neural Network, Decision Tree and Statistical Techniques Applied for Designing and Developing E-mail Classifier

H.S. Hota, Akhilesh Kumar Shrivas, S.K. Singhai

Abstract: ***Due to increased bandwidth and strong infrastructure available for accessing internet, internet users are growing rapidly. Internet users frequently use e-mail for fast data communication of audio, vedio and textual data but at the same time they are facing problem due to unwanted e-mail known as spam e-mail. In order to filter this unwanted e-mail, a classifier must be placed in the network or in computer. In this paper three different types of technique: Artificial Neural Network (ANN), Decision tree and statistical technique are explored for designing and developing e-mail classifier. Experimental work has been performed on e-mail data set obtained from UCI repository site and is partitioned into three different partitions to find out best suitable partition to be applied for various model. A suitable ensemble model is chosen based on various error measures calculated after training and testing the models. A final ensemble model is measured in terms of accuracy, precision, recall, F-measure and Gain Chart. Highest accuracy of 94.35% is obtained in case of ensemble of C5.0 and SVM with 60%-40% (training – testing) partition.***

Index Terms**: *C5.0, Support Vector Machine (SVM), Artificial Neural Network (ANN), Ensemble model.***

I. INTRODUCTION

Today business and academic needs has increased E-mail users and still it is increasing rapidly also E-mail is one of the most common, fast and economical means of communication over internet. One can send huge amount of text, audio and video data from one computer to another throughout the world within a fraction of second but this increased numbers of E-mail users are facing problem of spam E-mail. One out of every 12 e-mail received is a spam E-mail. Spam E-mail is accentually a junk E-mail send by spammers for their own bonafide intension. Spammers send the e-mail to attract e-mail users which belongs to various categories like making friend, lottery, adult news, advertisement etc. They collect e-mail addresses from the various sources like website, newsgroup, social website etc. and sends spam E-mail in bulk.

Revised Manuscript Received on 30 January 2013.
* Correspondence Author
Dr. H.S. Hota, Asst. Professor, Department of CS/IT, Guru Ghasidas University Bilaspur,(C.G.), India
Akhilesh Kumar Shrivas, Research Scholar, CVRU, Bilaspur (C.G.), India
Dr. S.K. Singhai, Department of Electronics Engineering, Govt. Engineering College, Bilaspur (C.G.), India

These huge number of spam e-mail are creating serious problem in terms of Communication bandwidth utilization, storage space in mailbox and time consumed to delete or maintain.

Due to these reasons spam filter must be associated to the network or computer which will separate spam and non spam mail before entering it into the mail box of any E-mail user. Many authors have done lot of work in this area using various data mining and soft computing techniques for image spam and text spam classification. A survey of machine learning techniques for spam filtering has recently published in which author has explored many data mining techniques like k-nearest neighbor, Artificial Neural Network, Naive bayes etc.[5], author(s) suggested that machine learning techniques may be one of the best techniques for anti-spam filtering. Decision tree is another data mining technique which is being used by many researchers for classification problem. A proposed model is based on this technique for spam e-mail classification, and achieved 94.6% accuracy which is higher than other individual decision tree based techniques like C4.5 [4].The same techniques is also applied by many other authors.

Ensemble of various data mining, soft computing and statistical techniques for spam e-mail classification problem is one of the major research area due to high accuracy achieved by many author. An ensemble model of Partical Swarm Optimization (PSO) and Support Vector Machine (SVM) was proposed by author's[6] ,similarly other author have integrated other techniques to form hybrid or ensemble model and tested the model on various spam e-mail related data set available publicly in repository data sites.In this paper an attempt has been made to develop an ensemble model based on ANN, Decision tree and statistical techniques. The developed ensemble models are compared with individual models in terms of various error measures like accuracy, precision, recall and F-measure. Classification result obtained in case of ensemble of C5.0 and SVM is satisfactory.

II. FRAMEWORK FOR CLASSIFICATION

A framework for classification of spam E-mail data is depicted in Figure 1. This framework can be viewed as four stages: Data Partition, Model Building, Model Validation detail and Performance Measure. All these phases are explained in more detail as below:

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Phase I: Data Partition

Spam E-mail data set used for training and testing of various models is downloaded from UCI repository site [10]. This data set is publicly available and can be downloaded freely. Data set contain 57 features related to E-mail with 4601 samples in all, out of which 1813(37.4%) samples are related to spam while rest of the samples i.e. 2788 (60.6%) are related to non-spam. In order to train and test models data set is divided into two parts: training and testing with 60-40%, 75-25% and 80-20% ratio. Both training and testing samples contain equal ratio of spam and non-spam related data.

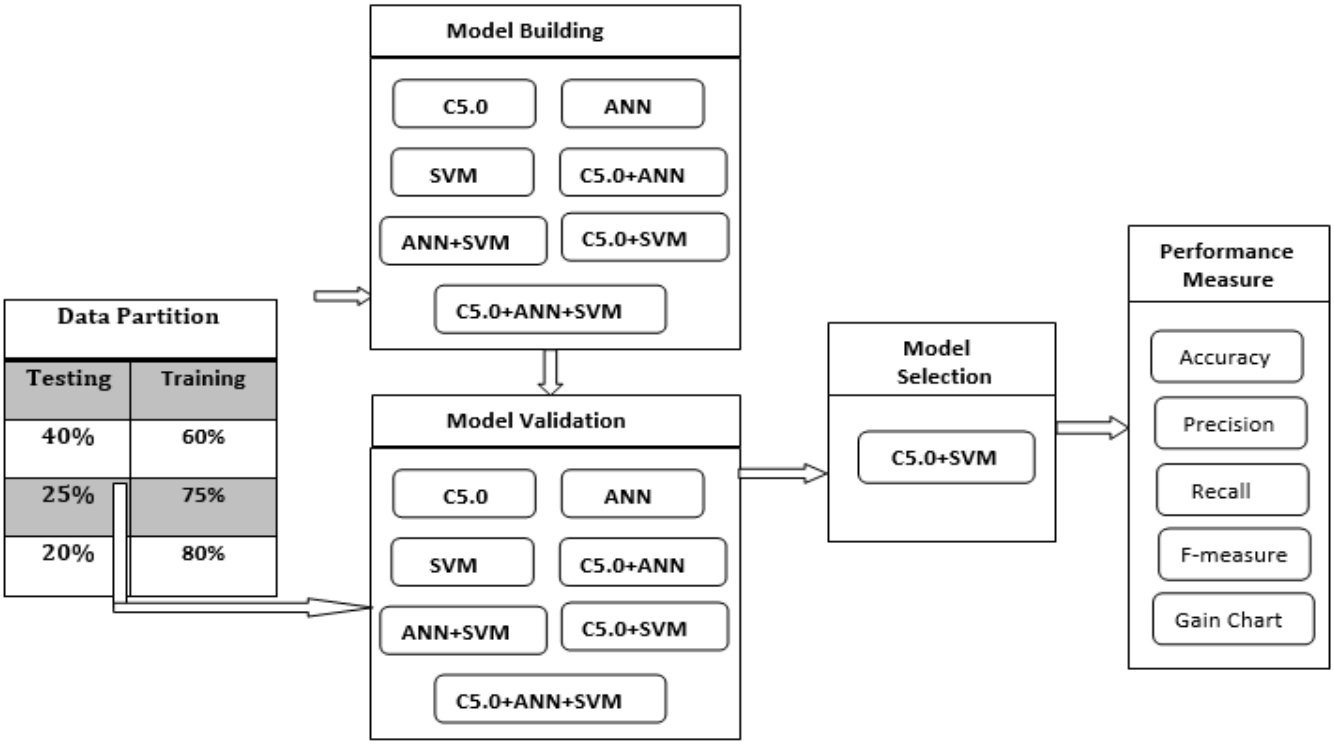


Fig 1: Framework for E-mail data classification

Phase II: Model Building

Model building is a phase where models are built by feeding training data set as partitioned in phase I. Three different categories of techniques are used to build model with all three partition of training data set to built classifier. This classification problem can be treated as binary classification problem, since there are only two classes that are to be classified by the classifier. Various techniques are explained in more detail as below:

A. Decision Tree (DT):

Decision tree [2] is most popular and powerful classification techniques in which in the training stage a tree like structure is formed where each non-leaf node is
decision node which splits according to the features of training data while leaf node represent class node, Once the decision tree is formed, unknown samples can be presented to the root node of decision tree and ultimately reaches to the class node to classify the sample as one of the target class.C5.0 is the latest development in this series and is an advance version of C4.5.This algorithm produces very high accuracy with lower execution time than any other decision tree based classification techniques.

B. Artificial Neural Network (ANN)

An Artificial neural network [2] is composed of a set of elementary computational units, called neurons, connected together through weighted connections. These units are organized in layers so that every neuron in a layer is exclusively connected to the neurons of the preceding layer and the subsequent layer. A multilayer feed-forward neural network consists of an input layer, one or more hidden layers, and an output layer.ANN is known as best classifier and is able to mine huge amount of data for classification.

There are many benefits of using ANN as classifier like simple architecture, high speed learning, able to input-output pattern matching .A popular ANN known as Error Back Propagation Neural Network (EBPN) is used here to develop E-mail classifier with one hidden layer.EBPN uses Error back propagation algorithm to train the network with two steps: forward phase in which input data are supplied and produce output is compared with actual output and finally error at outer layer and in Backward Phase: this error are sent back to the previous layer (hidden and input) to adjust the weight.

C. Support Vector Machine (SVM)

In formal definition, support vector machines [7] design a hyper planes or set of hyper planes in a high or infinite dimensional space, which can be used for classification, regression or other tasks.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

A SVM is a promising new method for classification of both linear and nonlinear data. SVM is 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 algorithms divide the n dimensional space representation of the data into two regions using a hyper plane. This hyper plane always maximizes the margin between the two regions or classes. The margin is defined by the longest distance between the examples of the two classes and is computed based on the distance between the closest instances of both classes to the margin, which are called supporting vectors.

D. Ensemble Model

An ensemble model is a combination of two or more models to avoid the drawbacks of individual models and to achieve high accuracy. The two models are combined by using high confidential wins scheme [1] where weights are weighted based on the confidence value of each prediction. Then the weights are summed and the value with highest total is again selected. The confidence for the final selection is the sum of the weights for the winning values divided by the number of models included in the ensemble model. If one model predicts no with a higher confidence than the two yes predictions combined, then no wins. We have formed various ensemble models and examined all these in terms of accuracy, following ensemble models are developed for further evaluations: Ensemble of C5.0 and ANN, ensemble of C5.0 and SVM, ensemble of ANN and SVM and ensemble of C5.0, ANN and SVM.All the above models have been trained and finally tested with training and testing email- spam data as shown in figure 1.

Phase III

All the trained models are testing using various partitions of testing data sets one by one and models are measured in terms of accuracy. A suitable model with highest accuracy is then selected for further evaluation.

Phase IV. PERFORMANCE MEASUREMENT

Performance of each individual classifier and its ensemble classifier can be evaluated by using some very well-known statistical measures: classification accuracy, precision, recall and F-measure. These measures are defined by true positive (TP), true negative (TN), false positive (FP) and false negative (FN).

Confusion matrix [2] for two classes is shown in table 1 where TP refers number of positive samples which is correctly classified by classifier, TN is number of negative samples classified correctly by the classifier, similarly FP are number of negative samples that is incorrectly classified (sample of class spam for which the classifier predicted non-spam) where as FN are the number of positive sampler that is incorrectly classified (sample of class non-spam for which classifier predicted spam).

Table 1: Confusion matrix for positive and negative

Samples

Predicted / Actual	Positive	Negative
Positive	True Positive	False Negative
Negative	False Positive (FP)	True Negative (TN)

Table 2: Various Performance Measures

Performance measure	Formula	Description
Classification Accuracy	(TP+TN)/N	Classification accuracy of classifier is the proportion of instances which are correctly classified.
Precision	TP/ (TP+FP)	Precision is the rate of instances classified correctly among the result of classifier.
Recall	TP/ (TP +FN)	Recall is the rate of correct classified instances among them to be classified correctly.
F-measure	$2 \times \frac{(\text{precision} \times \text{recall})}{(\text{precision} + \text{recall})}$	F-measure is the harmonic mean of precision and recall.

If the total number of cases are N then based on the table1, statistical performance measures can be evaluated using formula mentioned in table2.

Gain chart is another way to check the classifier which plots the values in the gains (%) column from the table. Gains are defined as the proportion of hits in each increment relative to the total number of hits in the tree, using the equation:

(Hits in increment / total number of hits) x 100%

Cumulative gains charts always start at 0% and end at 100% as we go from left to right. For a good model the gains chart will rise steeply toward 100% and then level off. A model that provides no information will follow the diagonal fro cumulative gains charts always start at 0% and end at 100% as we go from left to right. For a good model, the gains chart will rise steeply toward 100% and then level off. A model that provides no information will follow the diagonal from lower left to upper right.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

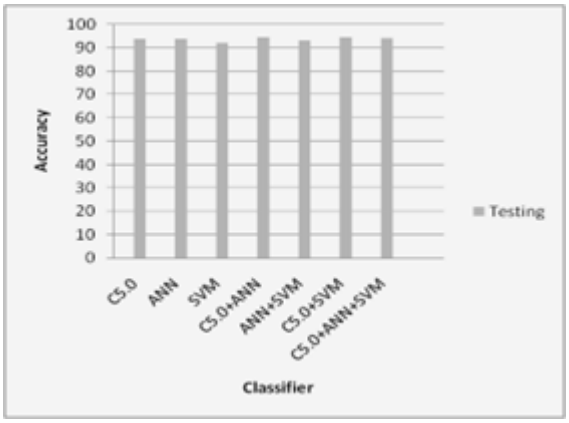
III. EXPERIMENT AND RESULTS

Experimental work is carried out using SPSS Clementine 12.0 software to design and develop classifier; objective of these classifiers is to distinguish between spam and non-spam with as much as higher accuracy. Experiment is done with three different partitions of data set as explained in phase 1 .Classification accuracy at testing stages for all the models are calculated as shown in table 3 using formula mention in table 2 after forming confusion matrix. Among the individual models, ANN is performing better than others for all three partitions but when ANN is ensemble with C5.0 or SVM then performance is increasing as it is 94.24% and 93.95% in case of ensemble of C5.0 and ANN and ensemble of SVM and ANN respectively for partition 60-40% .Similar trends are also followed in case of other two partitions, but when ANN is removed as one of the model of ensemble model and when SVM and C5.0 are combined to form a new ensemble model ,accuracy for 60-40% partition is achieved highest at 94.35% while it is decreased in case of other two partitions.

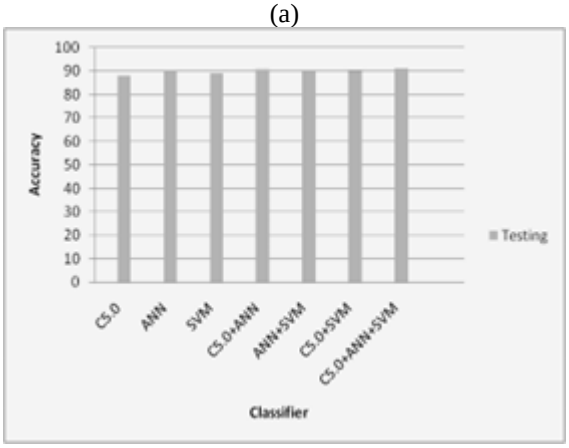
Accuracy obtained in this case is higher than that of work already done by [9], so there is a significant improvement of accuracy of individual as well as ensemble models due to suitable partition of E-mail data set. Other error measures like precession, recall and F-measure are calculated using equation shown in table 2 with the help of confusion matrix other performance measure like precision, recall and F-measure of the best model are 94.38%, 96.41% and 95.38% respectively for testing data set as shown in table 5.All the error measures of best ensemble model are also shown in form of bar graph in Figure 3. Gain chart is another way to check the performance of the model at testing stage which plots the value in gain% shown in Figure 4.This chart also proves that model is working well for classification of E-mail data set.

Table 3: Classification Accuracy of Various Models

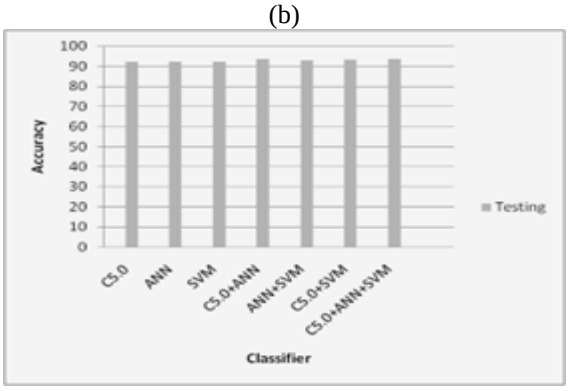
Models	Partition1 (60-40%)	Partition2 (75-25%)	Partition3 (80-20%)
	Testing	Testing	Testing
C5.0	93.75	92.35	87.83
ANN	93.86	92.35	89.67
SVM	91.91	92.26	89.13
Ensemble of C5.0 and ANN	94.24	93.57	90.65
Ensemble of ANN and SVM	93.95	93.04	90.22
Ensemble of C5.0 and SVM	**94.35**	93.3	90.11
Ensemble of C5.0,ANN and SVM	94.13	93.57	90.98



(a)



(b)



(c)

Fig 2: Testing accuracy for (a) Partition1 (b) Partition2 (c) Partition 3

Table 4: Confusion Matrix

Actual Vs. Predicted	Ensemble models of C5.0 and SVM	
	Non spam	Spam
Non Spam	1075 (TP)	40 (FN)
Spam	64 (FP)	662 (TN)

Table 5: Performance measure

Performance measures	Ensemble of C5.0 and SVM (Best Model)
Accuracy	94.35%
precision	94.38%
recall	96.41%
F-measure	95.38%

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

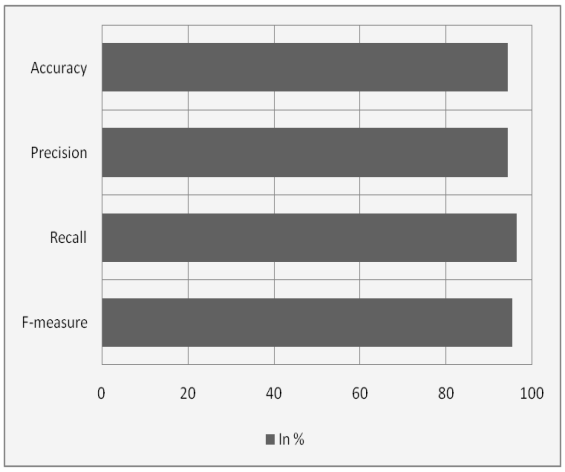


Fig 3: Error measures of best ensemble models

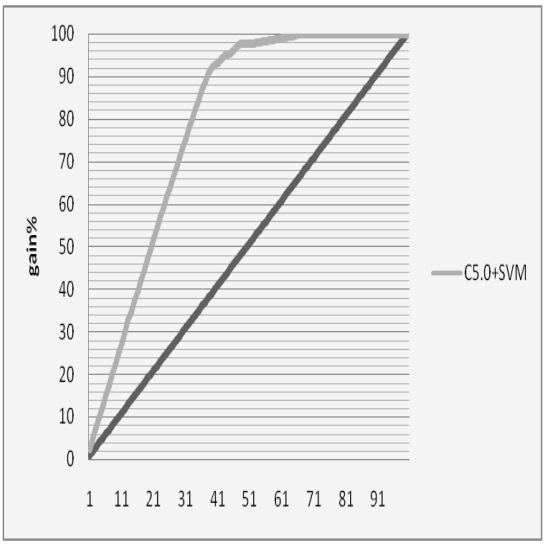


Fig 4: Gain chart for ensemble of C5.0 and SVM

V. CONCLUSION

Designing and developing a robust classifier for E-mail data classification in order to provide security to the E-mail users are challenging task and is a major research area. Authors are using many techniques to design and develop a suitable classifier for this purpose. In this paper three different categories of technique: Artificial Neural Network, Decision Tree and Statistical are applied and trained and tested with the help of three different partitions of data sets prepared, this work is an extended research work already done by the authors [9]. It is concluded that partition size plays an important role to improve classification accuracy. Highest classification accuracy is achieved in case of ensemble of SVM and C5.0 for 60-40% partition, due to ANN, models are (See the accuracy of ensemble model with ANN) producing highest accuracy but when it is combined with ensemble of SVM and C5.0 accuracy of model is decreased (94.13%) for partition 60-40% while it is increased for another two partitions (93.57% and 90.98%),this may be overtraining or under training characteristics of ANN. Other measures like precision, recall and F-measure are also calculated as 94.38%, 96.41% and 95.38% respectively which are satisfactory.

REFERENCES

1. Elsayad, A. M., 2010, “Predicting the severity of breast masses with ensemble of Bayesian classifiers”, Journal of Computer Science, 6(5), pp.576-584.
2. Jiawei Han and Micheline Kamber, 2006, “Data Mining Concepts and Techniques “, Morgan Kaufmann, San Francisco, Second Edition.
3. John Wang, 2003, “Data Mining: opportunities and challenges”, Idea Group, USA.
4. Lei SHI, Qiang WANG, Xinming MA, Mei WENG, Hongbo QIAO,2012, “Spam E-mail classification using Decision Tree Ensemble”, Journal of Computational Information Systems, pp. 949-956.
5. Omar Saad, Ashraf Darwish and Ramadan Faraj,2012,” A survey of machine learning techniques for Spam filtering”, IJCSNS International Journal of Computer Science and Network Security, vol.12 No.2.
6. R. Parimala, Dr. R. Nallaswamy, 2011,”A study of a spam E-mail classifications using feature selection package”, global journals Inc. (USA), vol.11.
7. V. N. Vapnik, 1998, “Statistical Learning Theory”, New York: John Wiley and Sons.
8. SPSS Clementine help file, 2007. http//www.spss.com.
9. Hota H.S. et al. ,2012,”Data mining techniques and its ensemble model applied for classification of e-mail data”, proceeding of review of business and technology research (RBTR) in International conference EPPICTM 2012,vol. 5 ,No. 1, ,pp. 473-479.
10. UCI Machine Learning Repository of machine learning databases (2010). University of California, school of Information and Computer Science, Irvine.C.A. http://www.ics.uci.edu/~mlram,?ML.Repositary.html

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)