Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/110419

An Ensembled Classifier for Email Spam Classification in Hadoop Environment

*D. Karthika Renuka**, *P. Visalakshiand SP. Rajamohana*

Department of Information Technology, PSG College of Technology, Coimbatore-641004, India.Department of ECE, PSG College of Technology, Coimbatore-641004, India.

Received: 3 Apr. 2017, Revised: 23 May 2017, Accepted: 26 May 2017
Published online: 1 Jul. 2017

Abstract: Email is one of the most ubiquitous and pervasive application used on a daily basis by millions of people worldwide. Email spam is a serious worldwide problem which causes problems for almost all computer users. Nowadays, e-mail becomes a powerful tool for communication as it saves a lot of time and cost. But, due to social networks and advertisers, most of the e-mails contain unwanted information called spam. Spam is the unwanted and unsolicited commercial e-mail. It is also known as junk e-mail. This issue not only affects normal users of the internet, but also causes a huge problem for companies and organizations since it costs a huge amount of money in lost productivity, wasting user's time and network bandwidth. Recently, various researchers have presented several email spam classification techniques. Spam classifications, which filter the spam emails from inbox moves it to our junk email folder. It automatically classifies email based on the social features. Spam classifies the set of mails into spam and ham based on its contents. It is very difficult to eliminate the spam mail completely as the spammers change their techniques frequently. The proposed system, we have developed is an efficient technique to classify the email spam using ensemble method. Gradient Boost classification is used which is an ensemble of the weak decision tree and weighted majority voting is used to ensemble the decision tree and also Naive Bayes classification is used. It consists of two phases, such as training phase and testing phase. The performance metrics namely precision, recall and accuracy are used for evaluation.

Keywords: Email spam, Hadoop, Map Reduce, Oracle VMVirtual Box, Apache Spark, Naive Bayes, Gradient Boost classification.

1 Introduction

Email spam, also known as junk email, is a type of electronic spam where unsolicited messages are sent by email. Many email spam messages are commercial in nature but may also contain disguised links that appear to be for familiar websites but in fact lead to phishing web sites or sites that are hosting malware. Spam email may also include malware as scripts or other executable file attachments. Spam is named after Spam luncheon meat by way of a Monty Python sketch in which Spam in the sketch is ubiquitous, unavoidable and repetitive. Email spam has steadily grown since the early 1990s. Botnets, networks of virus-infected computers, are used to send about 80% of spam. Since the expense of the spam is borne mostly by the recipient, it is effectively postage due advertising. The legal status of spam varies from one jurisdiction to another.

Spammers collect email addresses from chatrooms, websites, customer lists, newsgroups, and viruses that harvest users address books. These collected email addresses are sometimes also sold to other spammers. The proportion of spam email was around 80% of email messages sent, in the first half of 2010.

2 Related Work

Spam detection is a program used to detect spam e-mail and prevent those e-mails from entering into user's inbox [7], [13]. Recent research show that spam detection is usually processed by machine learning (ML) algorithms to distinguish between non-spam and spam e-mails. ML methods are able to extract the knowledge from a group of e-mails supplied and use the gained information in the categorization of newly received email. The aim of ML is to optimize the performance of the computer program through data or experience to make better decisions and solve problems in an intelligent way by using illustration data.

* Corresponding author e-mail: karthirenu@gmail.com

Jon Kagstrom [6] proposes random forest as a good solution to the scalability issues of single decision tree. Adboost algorithm is proposed by Youn, S. and D. McLeod is a meta-classification algorithm which can be combined with other classification algorithms to enhance its performance. Multi-class logistic classifier proposed in [7] is another important improvement of enhancing the basic boosting algorithm. K nearest neighbor [5] classifies samples based on the adjacent spatial relationships of features.

In the work by R. Geetha Ramani, G. Sivagami [15] employs supervised machine learning techniques namely C4.5 Decision tree classifier, Multilayer perceptron and Naive Bayes classifier. Five features of an e-mail: all (A), header (H), body (B), subject (S), and body with subject (B+S), are used to evaluate the performance of four machine learning algorithms. The training dataset, spam and legitimate message corpus is generated from the mails that they have received from their institute mail server for a period of six months. They conclude Multilayer Perceptron classifier outperforms other classifiers and the false positive rate is also very low compared to other algorithms.

There are some research work that apply machine learning methods in e-mail classification, M. N. Marsono, M. Watheq El-Kharashi, Fayez Gebali [1] They demonstrated that the naïve Bayes e-mail content classification could be adapted for layer-3 processing, without the need for reassembly. Suggestions on predetecting e-mail packets on spam control middleboxes to support timely spam detection at receiving e-mail servers were presented. M. N. Marsono, M. W. El Kharashi, and F. Gebali [1] They presented hardware architecture of naive Bayes inference engine for spam control using two class e-mail classification. That can classify more 117 millions features per second given a stream of probabilities as inputs. This work can be extended to investigate proactive spam handling schemes on receiving e-mail servers and spam throttling on network gateways. Savita Pundalik STeli and Santosh Kumar Biradar [2] proposed a system that used the SVM for classification purpose, such system extract email sender behavior data based on global sending distribution, analyze them and assign a value of trust to each IP address sending email message, the Experimental results show that the SVM classifier is effective, accurate and much faster than the Random Forests (RF) Classifier.

Tariq R. Jan et al. proposed a spam-based classification scheme of three categories. In addition to typical spam and not spam categories, a third undetermined category is provided to give more flexibility to the prediction algorithm. Undecided emails must be reexamined and collect further information to be able then to judge whether they are spam or not. Authors used Sculley and Cormack, 2008 and UCI Machine Learning Repository, as their experimental email dataset.

Izzat Alsmadi et al.'s (2015) paper evaluates applying rough set on spam detection with different rule execution schemes to find the best matching one. UCI Spam base is used in the experimental study. Instead machine learning approach uses, a set of training samples, these samples is a set of pre classified e-mail messages. Machine learning approach is more efficient than knowledge engineering approach; it does not require specifying any rules [4]. A specific algorithm is used that helps the machine to learn classification rules from these e-mail messages.

A Naive Bayes classifier applies Bayesian statistics with strong independence assumptions on the features that drive the classification process. Essentially, the presence or absence of a particular feature of a class is assumed to be unrelated to the presence or absence of any other feature. Bayesian spam filtering is a form of e-mail filtering that uses the naïve Bayesian classifier to identify spam e-mail [2]. Essentially, the presence or absence of a particular feature of a class is assumed to be unrelated to the presence or absence of any other feature. Bayesian spam filtering is a form of e-mail filtering that uses the naive Bayesian classifier to identify spam e-mail [2].

The main strength of naïve Bayes algorithm lies in its simplicity. Since the variables are mutually independent, only the variances of individual class variables need to be determined rather than handling the entire set of covariances. This makes naive Bayes one of the most efficient models for email filtering. It is robust, continuously improving its accuracy while adapting to each user's preferences when he/she identifies incorrect classifications thus allowing continuous rectified training of the model. In [3], the authors constructed a corpus Ling-Spam with 2411 non spam and 481 spam messages and used a parameter λ to induce greater penalty to false positives. They demonstrated that the weighed accuracy of a naïve-Bayesian email filter can exceed 99%. Variations of the basic algorithm for example, using word positions and multi-word N-grams as attributes have also yielded good results [4].

The first scholarly publication on Bayesian spam filtering was by Sharma, S., & Arora [12]. A naive Bayes classifier [3] simply apply Bayes theorem on the context classification of each email, with a strong assumption that the words included in the email are independent to each other. In the beginning, two sample emails from the real life data in order to create the training dataset.

Among all the different ways of implementing naive Bayes classification, Paul Graham's approach has become fairly famous [2]. He introduced a new formula for calculating token values and overall probabilities of an email being classified as spam. But there is a unrealistic assumption in his formula which assume the number of spam and ham are equal in the dataset for everyone. Where Tim Peter correctly adjust the formula later to make it fit into all datasets [7], and both methods will be evaluated using our training dataset.

Megha Rathi et al. (2013) In this paper the author exhibited the data mining techniques and also explained the classification algorithms. They evaluated various classification algorithms such as Naïve Bayes, Bayesian

Net, Random Forest, Random Tree, SVM etc. without feature selection first. Then they evaluated all these classification algorithms with feature selection by best first algorithm. The author analyzed that the Random Tree has 90.43% accuracy, which is very low. But with feature selection it reaches to 99.71% which is very high i.e. close to 100%. Therefore, they concluded that random tree is the best classification algorithm for email classification with feature selection.

Decision tree consist of the root node, branches and leaf nodes. In this, the tree is created in a top-down, recursive and divide and conquer way. It works like a greedy technique. The internal node defines the condition on the attribute, each branch defines the output of the condition and each leaf node defines the class label. [15]

The boosting algorithms are techniques to combine a number of weak learners to form an ensemble. The term weak leaner arrives from the PAC (probably approximately correct) [7], [8] learning community and indicates that the learning algorithm can learn with error rate slightly better than 50%. C4.5 classification trees are candidate weak learners even though their error rates can be much better than 50%. This version of boosting works as following: train the first member of the ensemble with training samples. In order to train the next member of the ensemble, the probability that a training sample will be picked to train the second member of the ensemble is adjusted upwards for hard examples and down for easy examples. By hard examples, we mean those examples that the first weak learner misclassifies. Each member of the ensemble is subsequently trained on examples picked from the original training set with their probabilities adjusted upwards or downwards depending on whether the previous members of the ensemble classified the training pattern incorrectly or correctly, respectively.

Finally, we note that the proposed approach is closely related to the family of ensemble approaches for semi-supervised learning. Ensemble methods have gained significant popularity under the realm of supervised classification, with the availability of algorithms such as Ada Boost. The semi-supervised counterparts of ensemble algorithms rely on the cluster assumption, and prime examples include ASSEMBLE and Semi-supervised Margin Boost (SSMB). Both these algorithms work by assigning a pseudo label to the unlabeled samples, and then sampling them for training a new supervised classifier.

3 Proposed Approach

3.1 Naive Bayes Classification

Naive Bayes-classifier is a simple probabilistic classifier based on applying Bayes theorem with strong independence assumptions. A more descriptive term for the underlying probability model would be independent feature model. The Naive-Bayes inducer computes conditional probabilities of the classes given the instance and picks the class with the highest posterior. One of the easiest ways of selecting the most probable hypothesis given the data that we have that we can use as our prior knowledge about the problem. Bayes Theorem provides a way that we can calculate the probability of a hypothesis given our prior knowledge. Bayes Theorem is stated as

$$P(h|d) = (P(d|h) * P(h)) / P(d)$$

Where P(h|d) is the probability of hypothesis h given the data d. This is called the posterior probability, P(d|h) is the probability of data d given that the hypothesis h was true, P(h) is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h, P(d) is the probability of the data (regardless of the hypothesis). After calculating the posterior probability for a number of different hypotheses, hypothesis with the highest probability is selected. This is the maximum probable hypothesis and may formally be called the maximum a posteriori (MAP) hypothesis. This can be written in any of the following way.

$$MAP(h) = max(P(hjd))$$
$$MAP(h) = max((P(djh) * P(h))/P(d))$$
$$MAP(h) = max(P(djh) * P(h))$$

where P(d) is a normalizing term which allows us to calculate the probability. We can drop it when we are interested in the most probable hypothesis as it is constant and only used to normalize. If we have an even number of instances in each class in our training data, then the probability of each class (e.g. P(h)) will be equal. Again, this would be a constant term in our equation and we could drop it so that we end up with

$$MAP(h) = max(P(d|h))$$

Naive Bayes-classifier is a probabilistic classifier based on conditional probability. It is a statistical technique. It is simple and easy to implement. Naive Bayes classifier exhibits high speed and accuracy when applied to large dataset. The probability for each mail to be ham or spam is calculated based on the Bayes theorem. The basic concept is to classify email as spam by looking at word frequency. The advantage is to improve the classification performance by removing the irrelevant features, good performance, it is short computational time. The disadvantage is the Naive Bayes classifier requires a very large number of records to obtain good results.

3.2 Gradient Boosting

Ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. Boosting is a sequential technique which works on

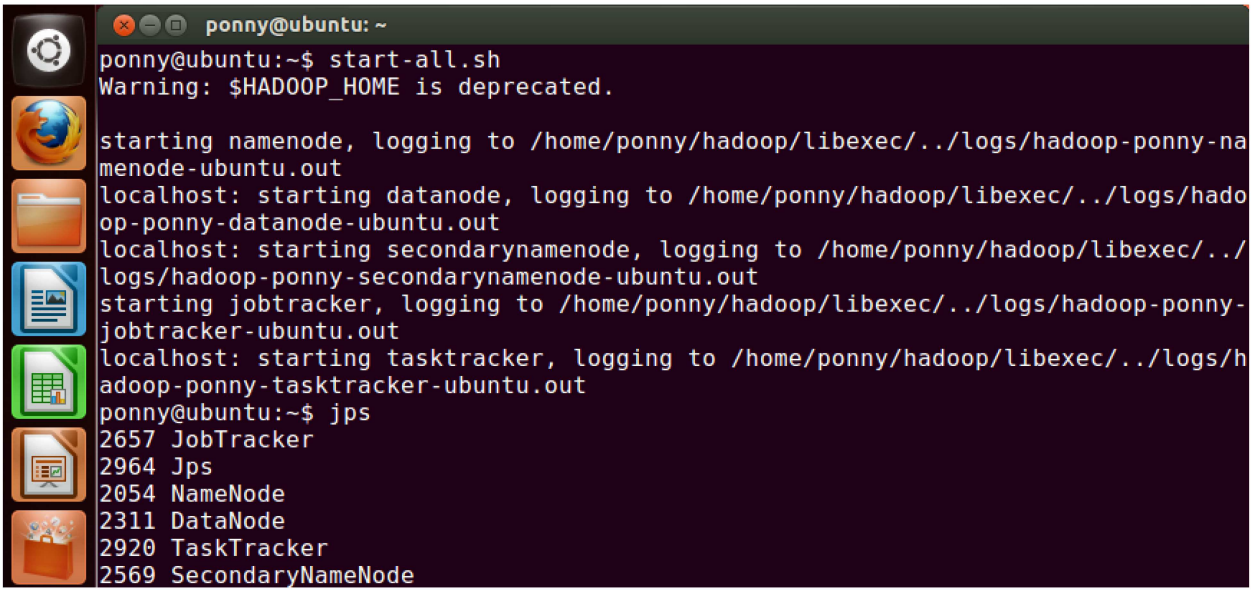

Fig. 1: Hadoop single node setup

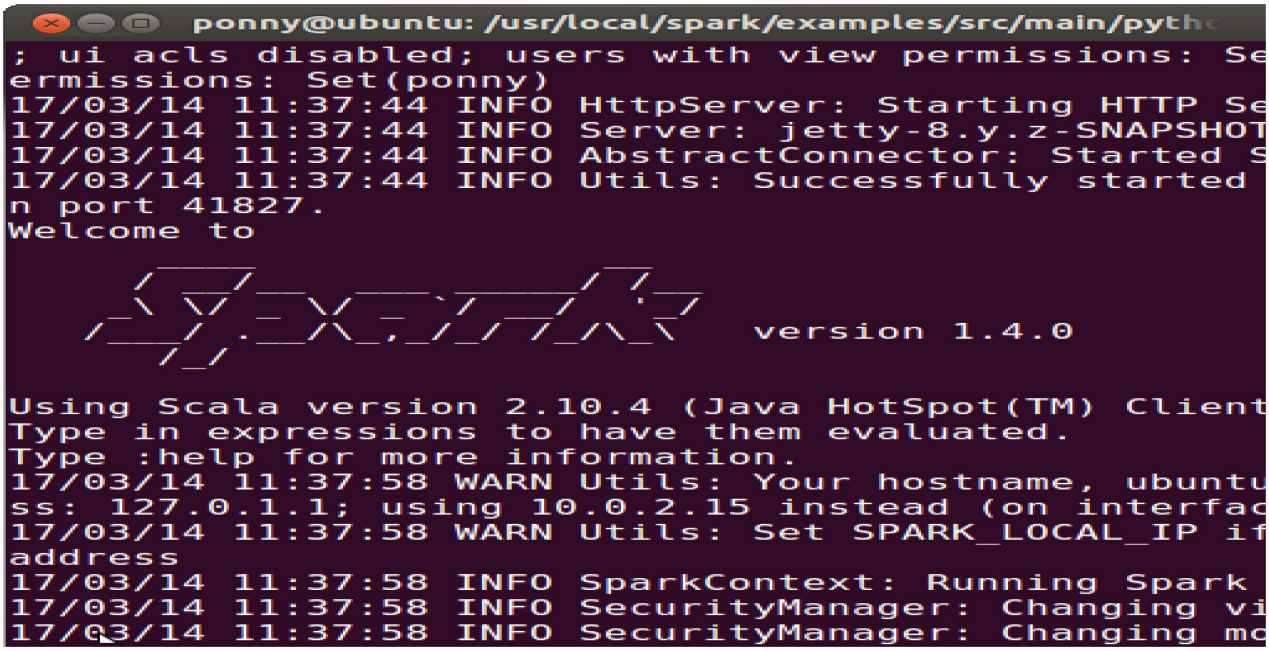

Fig. 2: Apache spark setup

the principle of ensemble. It combines a set of weak learners and delivers improved prediction accuracy. Boosted algorithms are used where we have plenty of data to make a prediction. And we seek exceptionally high predictive power. It is used to for reducing bias and variance in supervised learning. It combines multiple weak predictors to a build strong predictor. Boosting involves incrementally building an ensemble by training each new model instance to emphasize the training instances that previous models misclassified. In some cases, boosting has been shown to yield better accuracy than bagging, but it also tends to be more likely to over-fit the training data. In Boosting, each model is built on top of the previous ones. The final boosting ensemble uses weighted majority vote. Boosting can reduce the variance and bias of the base classifier.

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.Gradient boosting is able to provide smooth detailed predictions by combining many trees of very limited depth. Gradient boosting builds an ensemble of trees one-by-one, then the predictions of the individual trees are summed:

$$D(x) = dtree\ 1(x) + dtree\ 2(x) + \cdots$$

The next decision tree tries to cover the discrepancy between the target function $f(x)$f(x) and the current ensemble prediction by reconstructing the residual.

To get a bit closer to the destination, we train a tree to reconstruct the difference between the target function and the current predictions of an ensemble, which is called the residual.

$$R(x) = f(x) - D(x)$$

Spambase dataset is downloaded from UCI repository which consists of 4601 instances with 57 attribute characteristics of integer and real, multivariate dataset characteristics that are used for spam e-mail classification.

Table 1: Performance Measures

	NAIVE BAYES	GRADIENT BOOSTING
PRECISION	0.80	0.92
RECALL	0.78	0.89
F1 SCORE	0.81	0.91
ACCURACY	0.8077	0.941

A free and open-source hypervisor for x86 computers that supports the creation and management of guest virtual machines is installed. Above which hadoop single node setup is done successfully which is shown in the figure 1. After its successful installation spark framework is installed which is shown in the figure2. Apache spark is done. Spark is an open-source cluster-computing framework that supports Scala, R, Java and python. Naive bayes and Gradient boosting algorithm is implemented in the spark environment which is created.

The performance analysis is done for both naïve bayes and gradient boosting algorithm in the spark framework which is shown in the table1. The performance measures which is taken into account are precision, recall, f1 score and accuracy.

The following figure visualizes the performance measures of gradient boosting and naïve bayes in which gradient boosting shows the best results. Hence ensemble algorithm when compared is about to show the best result.

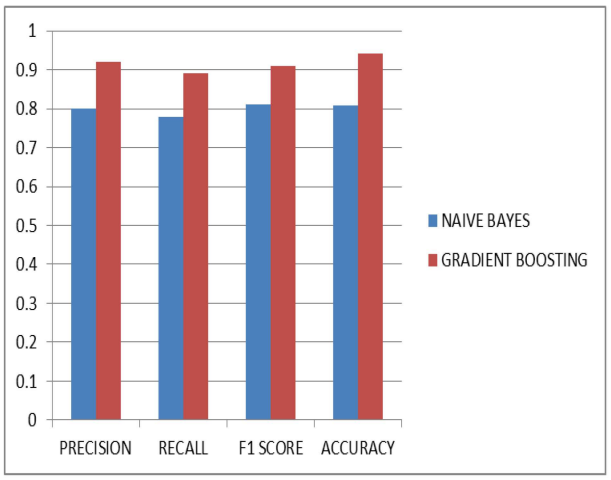

Fig. 3: Precision, recall, f1 score and accuracy values

4 Conclusion

This paper proposes a spam e-mail classification system using gradient Boost classification which is an ensemble of the weak decision tree and weighted majority voting is used to ensemble the decision tree and also naive bayes classification is used. The proposed spam e-mail classification system comprises of two phases, training and testing phases. It can be implemented in a distributed environment using single node hadoop environment to improve the performance of the classifiers. Further works have also been indicated in the following directions to facilitate the problem of e-mail spam classification. Different classifiers can also be ensembled to improve the performance of the classifier. The weight associated with each classifier in the ensemble method can be varied depending upon the performance of the classifier. To overcome the limitation of time complexity, the proposed algorithm can be implemented in a distributed environment using multi-node hadoop environment to improve the performance of the classifiers.

References

[1] M. N. Marsono, M. W. El-Kharashi, and F. Gebali, “Binary LNS-based naïve Bayes inference engine for spam control: Noise analysis and FPGA synthesis”, IET Computers & Digital Techniques, 2008

[2] Savita Pundalik STeli and Santosh Kumar Biradar, “Effective Email Classification for Spam and Non-spam,” International Journal of Advanced Research in Computer and software Engineering, vol. 4, pp. 273-278, 2014.

[3] Megha Rathi and Vikas Pareek, “Spam Email Detection through Data Mining-A Comparative Performance Analysis,” I.J. Modern Education and Computer Science, pp. 31-39, 2013.

[4] Youn, S. and D. McLeod, "A comparative study for email classification. Advances and Innovations in Systems", Computing Sciences and Software Engineering, 2007: p. 387-391.
[5] Izzat Alsmadi and Ikdam Alhami, "Clustering and Classification of email contents," Journal of King Saud and Information Sciences, pp. 46-57, 2015.
[6] Jon Kagstrom, "Improving Naive Bayesian Spam Filtering", M.Sc. Thesis, Mid/Sweden University Department for Information Technology and Media, spring 2005.
[7] Tariq R. Jan, "Effectiveness and Limitations of Statistical Spam Filters", University of Kashmir, Srinagar, India, International Conference on New Trends in Statistics and Optimization, 2009.
Jincheng Zhang, and Yan Liu, "Spam Email Detection: A Comparative Study", Techniques for Data Mining Journal, December 6, 2013.
[8] W.A. Awad, S.M. Elseuofi, "Machine learning methods for spam e-mail classification", International Journal of Computer Science & Information Technology, Vol. 3, No. 1, Feb 2011.
[9] Bhandarkar, M, "MapReduce Programming with Apache Hadoop", IEEE Conference on Parallel and Distributed Processing, April 2010.
[10] Olushola D. Adeniji, Olubukola Adigun, Omowumi O. Adeyemo3, "An intelligent spam scammer filter mechanism using Bayesian techniques", International Journal of Computer Science and Information Security (IJCSIS), Vol. 10, No. 3, March 2012.
[11] Sharma, S., & Arora, "Adaptive Approach for Spam Detection" International Journal of Computer Science Issues (IJCSI), 10(4), 2013.
[12] Xiao-li, C., Pei-yu, L., Zhen-fang, Z., & Ye, Q, "A method of Spam filtering based on weighted support vector machines." In IT in Medicine & Education, 2009.
[13] S. Nazirova, "Mechanism of classification of text spam messages collected in spam pattern bases," in Proceedings of the 3rd International Conference on Problems of Cybernetics and Informatics, (PCI '10), vol. 2, pp. 206–209, 2010.
[14] R. Geetha Ramani, G. Sivagami, "Parkinson Disease Classification using Data Mining Algorithms", International Journal of Computer Applications (0975 – 8887) Volume 32– No. 9, October 2011.
[15] https://archive.ics.uci.edu/ml/datasets/Spambase

Dhanaraj Karthika Renuka is working as Assistant Professor (SGr in the Department of IT, PSG College of Technology, Coimbatore, and TamilNadu, India. She obtained her master degree: ME (CSE) from Anna University Coimbatore. She obtained her Ph.D from Anna University, Chennai in the year 2015 .Her specializations include Evolutionary Computing, Soft Computing, Computer Networks, Information Security, Data Mining, and Deep Learning.

P. Visalakshi is a Masters in Applied Electronics and a Ph.D. in Information and Communication engineering from Anna University, Chennai. She has completed her Honors Diploma in Network Centered Computing from NIIT in the year 1998. She has over 15 years of teaching experience of which 13 years is in the Computer Science and Engineering Department in the same college. She currently works as an Associate Professor at ECE Department, PSG College of Technology. She has organized one International Conference and two National Level Conferences. She has been the CSE department Alumni coordinator from 2007-2012. She has been the coordinator for five National Level Technical Symposiums (Kruzade'08-Kruzade'12) and two technical workshops. She has been the faculty advisor of the Computer Science and Engineering Association from 2010-2012. She has been the coordinator for the PSG-NIC Projects. She has been the executive council member of the PSG Tech Alumni Association and currently the secretary of the PSG Tech Alumni Association.

SP. Rajamohana is working as Assistant Professor (Sr.Gr) in the Department of IT, PSG College of Technology, Coimbatore, and TamilNadu, India. She obtained her Bachelor's degree from Thiagarajar College of Engineering in 2006. She received her Master degree from PSG College of Technology in 2008. She is currently doing research in the area of Review spam detection. Her research areas include Data mining, Evolutionary Computation, Software Engineering.