Global Journals LaTeX JournalKaleidoscope™

Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

Comparative Study of Gaussian and Nearest Mean Classifiers for Filtering Spam E-mails

Dr. Upasna Attri[1] and Harpreet Kaur[2]

[1] Punjab Technical University, Jalandhar (India)

Received: 9 December 2011 Accepted: 3 January 2012 Published: 15 January 2012

Abstract

The development of data-mining applications such as classification and clustering has shown the need for machine learning algorithms to be applied to large scale data. The article gives an overview of some of the most popular machine learning methods (Gaussian and Nearest Mean) and of their applicability to the problem of spam e-mail filtering. The aim of this paper is to compare and investigate the effectiveness of classifiers for filtering spam e-mails using different matrices. Since spam is increasingly becoming difficult to detect, so these automated techniques will help in saving lot of time and resources required to handle e-mail messages.

Index terms— Data-mining, Machine Learning, Classifiers, Filtering, spam E-mails.

1 Introduction

he Internet is a global system of interconnected computer networks to serve billions of users worldwide. As of 2011, more than 2.1 billion people -nearly a third of Earth's population -use the services of the Internet. E-mail has become one of the fastest and most economical forms of communication due to minimal costs, reliability, accessibility and speed. Wide usage of e-mail prone to spam e-mails. Spam e-mail is junk or unwanted bulk e-mail or commercial e-mail for recipients. Various problems that exist from spam emails are: wastage of network time and resources, damage to computers and laptops due to viruses and the ethical issues like advertising immoral and offensive sites that are harmful to the young generations. It hardly cost spammers to send out millions of e-mails than to send few e-mails, causing financial damage to companies and annoying individual users. Spam filter software can help mitigate this overwhelming chore. No spam filter software is 100% effective. Spam mail can contain viruses, keyloggers, phishing attacks and more. Clearly, a war is waging inside a user's inbox. Deployments of better ways to filter spam e-mails are needed. Several major kinds of classification method including decision tree induction, Bayesian networks, knearest neighbor classifier, case-based reasoning, genetic algorithm, fuzzy logic techniques, Neural Network (NN), Support Vector Machine (SVM), and Naïve Bayesian (NB) are showing a good classification result. Among the approaches developed to stop spam, filtering is an important and popular one.

Author ? : CSE Department, DAV Institute of Engineering and Technology, Jalandhar, India. E-mail : upasnaa.08@gmail.com Author ? : CSE Department, DAV Institute of Engineering and Technology, Jalandhar, India. E-mail : harpreet_daviet@yahoo.in Recently, there is a growing emphasis on investigative analysis of datasets to discover useful patterns, called data mining. Data Mining is the extraction of interesting, valid, novel, actionable and understandable information or patterns from large databases for making decisive business decisions. Classification is a data mining (machine learning) technique used to predict group membership for data instances. Filtering is very important and popular approach to circumvent this problem of spam. For filtering spam e-mails from good ones, clustering technique is imposed as classification method on a finite set of objects. Clustering is the technique used for data reduction. It divides the data into groups based on pattern similarities such that each group is abstracted by one or more representatives.

Classification is a supervised learning method. The aim of classification is to create a model that can predict the 'type' or some category for a data instance that doesn't have one. There are two phases in classification: first is supervision in which the training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations. Second is prediction in which given an unlabelled, unseen instance, use the model to predict the class label. Some algorithms predict only a binary split (yes/no), some can predict 1 of N classes, and some give probabilities for each of N classes.

Clustering is an unsupervised learning. It is a method by which a large set of data is grouped into clusters of smaller sets of similar data. There are two phases in this method: In first phase the class labels of training data is unknown. Whereas in second phase, given a set of measurements, observations, etc. the aim is to establish the existence of classes or clusters in the data. There are no predefined classes. Besides the term clustering, there are a number of terms with similar meanings, including automatic classification, numerical taxonomy, botryology and typological analysis.

Various criteria to evaluate the best spam filter software as following:(D D D D)

In the knowledge engineering approach, a set of rules is created according to which messages are categorized as spam or legitimate mail. The major drawback of this method is that the set of rules must be constantly updated, and maintaining it is not convenient for most users. In the machine learning approach, it does not require specifying any rules explicitly. Instead, a set of pre-classified documents (training samples) is needed. A specific algorithm is then used to "learn" the classification rules from this data. The subject of machine learning has been widely studied and there are lots of algorithms suitable for this task.

Some of the existing approaches to solve the problem of spam mails could be listed as follows:II.

Statement of the Problem E-mail has been an efficient and popular communication mechanism as the number of Internet users increase. Therefore, e-mail management is an important and growing problem for individuals and organizations because it is prone to misuse. The blind posting of unsolicited e-mail messages, known as spam, is an example of misuse. Automatic e-mail filtering seems to be the most effective method for countering spam at the moment and a tight competition between spammers and spam-filtering methods is going on: the finer the anti-spam methods get, so do the tricks of the spammers. So, uses of machine learning algorithms are imposed to overcome this problem upto large extent. There is substantial amount of research is going on with machine learning algorithms. It works by first learning from the past data available for training and then used to filter the spam e-mails effectively. In this work, comparison of two machine learning algorithms is conducted. Gaussian and Nearest Mean classifiers are one of the most effective machine learning algorithms. Therefore, Comparison of these two algorithms is proposed to be conducted for investigating the effectiveness to filter the spam e-mails.

2 III.

3 Objective of Work

The goals of this paper are three fold.

4 Related Work

In this technical report (Sahami et al. 1998) developed probabilistic learning methods for filtering spam e-mail using Bayesian network. (Drucker et al. 1999) compared Support Vector Machine (SVM) with Ripper, Rochio and Boosting Decision Tree (classification algorithms) and concluded that Boosting Trees and SVMs had an acceptable performance in terms of accuracy and speed. In his paper (Tretyakov, 2004) ? Rules: should give the user the ability to edit predefined rule settings as well as the creation of new rules.

? Compatibility: compatible with their current e-mail client or web-mail service provider.

There are two general approaches to mail filtering:? Knowledge Engineering (KE) ? Machine Learning (ML).

? Rule based: Hand made rules for detection of spam made by experts (needs domain experts & constant updating of rules).

? Customer Revolt: Forcing companies not to publicize personal e-mail ids given to them (hard to implement).

? Domain filters: Allowing mails from specific domains only (hard job of keeping track of domains that are valid for a user).

? Blacklisting: Blacklist filters use databases of known abusers, and also filters unknown addresses (constant updating of the data bases would be required).

? White list Filters: Mailer programs learn all contacts of a user and let mail from those contacts through directly (every one should first be needed to communicate his e-mail-id to the user and only then he can send e-mail).

? Hiding address: Hiding ones original address from the spammers by allowing all e-mails to be received at temporary e-mail-id which is then forwarded to the original e-mail if found valid by the user (hard job of maintaining couple of e-mailids).

? Government actions: Laws implemented by government against spammers (hard to implement laws).

? Automated recognition of Spam: Uses machine learning algorithms by first learning from the past data available (seems to be the best at current).

? Checks on number of recipients:by the e-mail agent programs.

these algorithms achieve better precision as compared to each other.

In their work (Aery et al. 2005) concluded that structure and content of e-mails in a folder classifies effectively the incoming e-mails. (Kulkarni et al. 2005) in their paper concluded that e-mail messages can be treated as contexts and clustering is based on underlying content rather than occurrence of some specific string. In this technical report (Segal et al. 2005) presented SpamGuru: an anti-spam filtering system for enterprises that is based on three principles: plug-in tokenizers and parsers, plug-in classification modules and machine learning techniques. SpamGuru produces excellent spam detection results. In his work (Zhao C. 2005) combined three classifiers (k-NN, Classical Gaussian and Boosting with Multi-Layer Perceptron) to produce Mixture of Expert (MOE) and concluded that Boosting is effective and also outperforms MOE.

In their journal (Bratko et al. 2006) concluded that compression models outperform currently established spam filters. The nature of the model allows them to be employed as probabilistic text classifiers based on character-level or binary sequences. In his paper (Hoanca B. 2006) concluded that no e-mail control technique is 100% effective. This problem of spam is shifting to other communication medias also in the form of Spam on Instant Messages (SPIM) and in chat rooms (SPAT).

In this journal (Blanzieri et al. 2007) concluded that the feel of antispam protection in by now matured and well developed. But inboxes are full of spam. So, more sophisticated techniques and methods are required to mitigate this problem of spamming. In his paper (Lai C.C. 2007) compared three method (SVM, Naïve-Bayesian (NB) and k-NN) and concluded that NB and SVM outperforms k-NN using header of e-mails only. In their technical report (Youn et al. 2007) compared four classifiers (neural network, SVM, Naïve-Bayesian and J48) and concluded that J48 classifier can provide better classification results for spam e-mail filtering.

In this technical report (Blanzieri et al. 2008) concluded that now situation of spam is tolerable and one can give attention to produce robust classification algorithm. In this report (Sculley et al. 2008) showed the impact of noisy labeling feedback on current spam filtering methods and observed that these noise tolerant filters would not necessarily have achieved best performance.

In this journal (Xiao-Li et al. 2009) proposed spam detection using clustering, random forests and active learning with respect to term frequency and inverse document frequency for messages. (DeBarr et al. 2009) compared six classifiers to treat Arabic, English and mixed e-mails and concluded that features selection technique can achieve better performance than filters that do not used them. El-Halees A. (**??**009) proposed a semi supervised approach for image filtering and concluded that this approach achieves high detection rate with significantly reducing labeling cost. (Gao et al. 2009) discussed one of key challenges that effect the system which is identifying spammers and also discussed on potential features that describes system's users and illustrate how one can use those features in order to determine potential spamming users through various machine learning models has been done. These proposed features demonstrate improved results as compared to the previous work done on it. In their work (Madkour et al. 2009) improved NB classifiers and concluded better detection rate of precision when compared with some best variants of NB. (Song et al. 2009) When used into spam filtering, the standard support vector machine involves the minimization of the error function and the accuracy of the SVM is very high, but the degree of misclassification of legitimate e-mails is high. In order to solve that problem, a method of spam filtering based on weighted support vector machines. Experimental results show that the algorithm can enhance the filtering performance effectively.

In this paper (Basavraju et al. 2010) proposed a spam detection technique using text clustering based on vector space model and concluded that k-means works well for smaller data sets and BIRCH with k-NN in combination performs better with large data sets. In this paper (Gao et al. 2010) presented a comprehensive solution to image spam filtering which combine cluster analysis of spam images on server side and active learning classification on client side for effectively filtering image spam. In this journal (Nagwani et al. 2010) proposed a weighted e-mail attribute similarity based model for more accurate clustering.

V.

5 Materials and Methods

The Matlab has been used as the programming tool for this simulation experiment. Random samples for each class of e-mail were generated and random partitioning of the samples of each class into two equal sized sets to form a training set and a test set for each class has been done. For each case, estimated the parameters of the Normal density function from the training set of the corresponding class. For each case the estimates of the parameters have been used to determine the Gaussian discriminant function. The Gaussian classifier for spam problem has been developed. The test samples have been classified for each class. For each case, the probability of classification error (POE) has been determined and also the time taken (in seconds) to classify has been measured. Further the nearest mean classifier has been implemented. The test samples of each class have been classified. For each case, the probability of classification error (POE) has been estimated and also the time taken (in seconds) for classification has been measured.(D D D D)

Finally comparison of the two methods for effectiveness against spam based on probability of error and time taken to classify has been conducted.

6 VI.

7 Results and Discussions

During first execution 50 e-mail messages were generated and classified according to Gaussian and Nearest Mean method. The plot shows the variation of probability of error. It can be seen that the maximum POE is almost

0.108 in the case of Nearest Mean method and mostly the POE of the Gaussian method is generally less than the Nearest Mean method. However at some instances the POE of Gaussian method is more is at the 04 th and 15 th e-mail message (Fig. ??). Fig. ?? : The variation of probability of error for 50 E-mails When 100 e-mail messages were generated and classified according to Gaussian and Nearest Mean method then plot shows the variation of probability of error. It can be seen that the maximum POE is almost 0.087 in the case of Nearest Mean method and mostly the POE of the Gaussian method is generally less than the Nearest Mean method. However at some instances the POE of Gaussian method is more is at the 38 th and 76 th e-mail message (Fig. ??).

8 Fig. 2 : The variation of probability of error for 100 E-mails

When 150 e-mail messages were generated and classified according to Gaussian and Nearest Mean method, then plot shows the variation of probability of error. It can be seen that the maximum POE is almost 0.114 in the case of Nearest Mean method and mostly the POE of the Gaussian method is generally less than the Nearest Mean method. However at some instances the POE of Gaussian method is more is at the 40 th and 140 th e-mail message (Fig. ??). In the next iteration 250 e-mail messages were generated and classified according to Gaussian and Nearest Mean method. The plot shows the variation of probability of error. It can be seen that the maximum POE is almost 0.117 in the case of Nearest Mean method and mostly the POE of the Gaussian method is generally less than the Nearest Mean method. However at some instances the POE of Gaussian method is more is at the 120 th and 240 th e-mail message (Fig. ??).

9 June

During fourth execution 200 e-mail messages were generated and classified according to Gaussian and Nearest Mean method. The plot shows the variation of probability of error. It can be seen that the maximum POE is almost 0.104 in the case of Nearest Mean method and mostly the POE of the Gaussian method is generally less than the Nearest Mean method. However at some instances the POE of Gaussian method is more is at the 35 th and 109 th e-mail message (Fig. ??).

10 Comparison and Analysis

It is analyzed from the above results that most of the times Gaussian Classifier performs better (POE is less) than the Nearest Mean Classifier. But still there are few traces of Nearest Mean Classifier showing less POE than Gaussian Classifier (rare cases). To check the overall performance of these two methods, their average of POE is estimated as shown in Table ??

11 June

In the next experiment e-mail messages were generated and classified according to Gaussian and Nearest Mean method and the time taken to classify was plotted (Fig. ??). The plot shows that as the load of incoming e-mails increases the Gaussian classifier takes more time than the Nearest Mean classifier.

12 Sr

13 Conclusion

It can be seen from Fig- ?? to Fig- ?? that most of the times Gaussian method gives better performance and the POE is less as compared to Nearest Mean method. Still a few times the Nearest Mean method resulted in less POE but these instances are rare. But Table-1 shows that the average Probability of error (POE) of Gaussian Classifier is less (better) than that of Nearest Mean Classifier. From Fig- ?? it can be seen that as the load of incoming e-mails increases the Gaussian classifier takes more time than the Nearest Mean classifier. Table-2 shows that the average time taken by Gaussian classifier is more than the Nearest Mean classifier.

Since in filtering spam e-mails, more weightage is given to accuracy than the time taken to classify. So, it can be concluded that in filtering spam emails the method of Gaussian Classification is better than the Nearest Mean method.

IX.

Figure 1:

1

Sr. No.	No. of E-mails	POE (Avg) (Gaussian)	1 POE (Avg) (Nearest Mean)
1.	50	0.04587	0.05200
2.	100	0.04713	0.05107
3.	150	0.04876	0.05222
4.	200	0.04577	0.04933
5.	250	0.04680	0.05077

Figure 2: Table 1 :

2

	No. of E-mails	Avg. Time (in sec) Gaussian	Avg. Time (in sec) Nearest Mean
1.	100	0.04524	0.00717
2.	200	0.04391	0.00585
3.	300	0.04373	0.00577
4.	400	0.04403	0.00577
5.	500	0.04449	0.00577
6.	600	0.04368	0.00579
7.	700	0.04372	0.00577
8.	800	0.04408	0.00579
9.	900	0.04376	0.00577
10.	1000	0.04386	0.00578

Figure 3: Table 2 :

.1 Acknowledgment

The authors are grateful to DAV Institute of Engineering and Technology, Jalandhar (India) for providing continuous support throughout the work.

[Sahami et al. ()] *A Bayesian Approach to Filtering Junke-mail*, M Sahami , S Dumais , D Heckermaan , E Horvitz . 1998. Redmond WA. Stanford University and Microsoft Research (Technical Report)

[Youn and Mcleod ()] *A Comparative Study for E-mail Classification*, S Youn , D Mcleod . 2007. Los Angeles, CA 90089 USA. University of Southern California (Technical Report)

[Gao et al. ()] 'A Comprehensive Approach to Image Spam Detection: From Server to Client Solution'. Y Gao , A Choudhary , G Hua . *IEEE Transactions on Information Forensics and Security* 2010. 5 (4) p. .

[Xiao-Li et al. ()] *A method of Spam filtering based on weighted support vector machines*, C Xiao-Li , L Pei-Yu , Z Zhen-Fang , Q Ye . IEEE Xplore 978-1-4244- 3930-0/09. 2009.

[Basavraju and Prabhakar ()] 'A Novel Method of Spam Mail Detection using Text Based Clustering Approach'. M Basavraju , R Prabhakar . *International Journal of Computer Applications* 2010. 5 (4) p. .

[Blanzieri and Bryl ()] *A Survey of Learning-Based Techniques of E-mail Spam Filtering*, E Blanzieri , A Bryl . 2007. Italy. University of Toreto (Technical Report)

[Blanzieri and Bryl ()] *A Survey of Learning-Based Techniques of E-mail Spam Filtering*, E Blanzieri , A Bryl . 2008.

[Lai ()] 'An empirical study of three machine learning methods for spam filtering'. C C Lai . *ScienceDirect, Knowledge-Based Systems* 2007. 20 p. .

[Nagwani and Bhansali ()] 'An Object Oriented E-mail Clustering Model Using Weighted Similarities between E-mails Attributes'. N K ; Nagwani , A Bhansali . *International Journal of Research and Reviews in Computer Science (IJRRCS)* 2010. 1 (2) p. .

[Song et al. ()] 'Better Naïve Bayes classification for high-precision spam detection'. Y Song , A Ko?cz , C L Giles . *Softw. Pract. Exper. No* 2009. 39 p. .

[Aery and Chakravarthy ()] *e-mailSift: E-mail Classification Based on Structure and Content*, M Aery , S Chakravarthy . 2005. (supported by NSF grants IIS-0123730, IIS-0097517, IIS-0326505, and EIA-0216500)

[Sculley and Cormack ()] *Filtering E-mail Spam in the Presence of Noisy User Feedback*, D Sculley , G V Cormack . 2008. Medford, MA 02155 USA. Tufts University 161 College Ave (Technical Report)

[El-Halees ()] 'Filtering Spam E-Mail from Mixed Arabic and English Messages: A Comparison of Machine Learning Techniques'. A El-Halees . *The International Arab Journal of Information Technology* 2009. 6 (1) p. .

[Hoanca ()] 'How good are our weapons in spam wars?'. B Hoanca . *Technology and Society Magazine* 2006. 25 (1) p. . (IEEE)

[Kulkarni and Pederson ()] 'Name Discrimination and E-mail Clustering using unsupervised Clustering and Labeling of Similar context'. A Kulkarni , T Pederson . *2 nd Indian International Conference on Artificial Intelligence-(IICAI-05*, 2005. p. .

[Gao et al. ()] 'Semi Supervised Image Spam Hunter: A Regularized Discriminant EM Approach'. Y ; Gao , M Yang , A Choudhary . *ADMA 2009, LNAI*, 2009. 5678 p. .

[Debarr and Wechsler ()] 'Spam Detection using Clustering, Random Forests, and Active Learning'. D Debarr , H Wechsler . *CEAS 2009 -Sixth Conference on E-mail and Anti-Spam*, 2009.

[Bratko et al. ()] 'Spam Filtering Using Statistical Data Compression Models'. A Bratko , G V Cormack , T R Lynam . *Journal of Machine Learning Research* 2006. 7 p. .

[Segal et al. ()] *SpamGuru: An Enterprise Anti-Spam Filtering System*, R Segal , J Crawford , J Kephart , B Leiba . 2005. IBM Thomas J. Watson Research Center (Technical Report)

[Drucker et al. ()] 'Support vector machines for spam categorization'. H Drucker , W Donghui , V N Vapnil . *IEEE Transactions* 1999. 10 (5) p. .

[Zhao ()] *Towards better accuracy for Spam predictions*, C Zhao . 2005. Ontario, Canada. University of Toronto (Technical Report)

[Tretyakov ()] K Tretyakov . MTAT.03.177. *Machine Learning Techniques in Spam Filtering*, 2004. p. . (Data Mining Problem-oriented Seminar)

[Madkour et al. ()] 'Using Semantic Features to Detect Spamming in Social Bookmarking Systems'. A Madkour , T Hefni , A Hefny , K S Refaat . *Human Language Technologies Group IBM Cairo Technology Development Center, El-Ahram* 2009.