A COMPARATIVE STUDY OF CLASSIFIERS FOR FILTERING SPAM EMAILS

Upasna Attri[*1], Dinesh Kumar[2], Bhupinder Singh[1], Neha Kapur[1], Puneet Kaur[1]

[1]*C.S.E. Deptt, I.G.C.E. Abhipur, Mohali, India*

[2]*I.T. Deptt, D.A.V.I.E.T. Jalandhar, India*

***Abstract*— this paper presents the comparison between Gaussian classifiers and Nearest Neighbor Classifiers for filtering spam emails. The results are in the form of traces of probability of error and time taken for classification, both with respect to the number of emails. Since spam emails are increasingly becoming difficult to filter, so these automated techniques will help in saving lot of time and resources required to handle the same.**

***Keywords*—Knowledge discovery, Data mining, Spam emails**

I. Introduction

The knowledge discovery and data mining (KDD) field draws on findings from statistics, databases, and artificial intelligence to construct tools that let users gain insight from massive data sets. People in business, science [1], medicine [2], [3], [4], academia, and government collect such data sets, and several commercial packages now offer general purpose KDD tools. An important KDD goal is to “turn data into knowledge.” For example, knowledge acquired through such methods on a medical database could be published in a medical journal. Knowledge acquired from analyzing a financial or marketing database could revise business practice and influence a management school’s curriculum. In addition, some US laws require reasons for rejecting a loan application, which knowledge from the KDD could provide. Occasionally, however, you must explain the learned decision criteria to a court, as in the recent lawsuit Blue Mountain filed against Microsoft for a mail filter that classified electronic greeting cards as spam mail [5]. In one early KDD success story, Robert Evans and Doug Fisher analyzed data from a printing press, found conditions under which the press failed, and identified rules to avoid these failures [6], [7]. The increase in the affordability of storage capacity, the associated growth in the volumes of data being stored and the mounting recognition in the value of temporal data (as well as the usefulness of temporal databases and temporal data model modeling) has resulted in the prospect of mining temporal rules from both static and longitudinal data. Data mining can itself be viewed as the application of artificial intelligence and statistical techniques to the increasing quantities of data held in large, more or less structured data sets, such as databases and temporal data mining is an extension of this work [8]. The knowledge discovery process is comprised of business understanding, identifying data requirements, data preparation, modeling, evaluation, and deployment [9], [10], [11]. Many methods have been proposed to increase the efficiency of data mining algorithms [12], [13]. Although a lot of applications in business, science [14], [15], medicine [16], [17] have been developed but not many applications have been exploited to control spamming in internet. A few attempts made so far are experimental which take a lot of time and are expensive to conduct [18],[19].

II. Materials and Methods

The Matlab has been used as the programming tool for this simulation experiment. Random samples for each class of email were generated and random partitioning of the samples of each class into two equally 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 Gaussian 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 email problem has been developed. The test samples have been classified for each class. For each case, the probability of classification error has been determined and also the time taken to classify has been measured. Further the nearest neighbor 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 for classification has been measured. Finally comparison of the two methods for effectiveness against spam emails based on probability of error and time taken to classify has been conducted.

III. Results and discussion

In the first iteration 50 email messages were generated and classified according to Gaussian and nearest neighbor method. The plot shows the variation of probability of error. It can be seen that the maximum POE is almost 0.093 in the case of nearest neighbor method and mostly the POE of the Gaussian classification method is generally less than the nearest neighbor classification method. However at some instances the POE of Gaussian classification method is more at the 25th and 35th email message (Fig. 1).

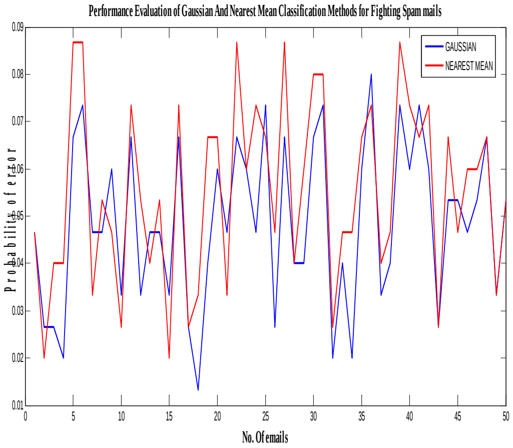

Fig. 1 It shows the variation of probability of error for 50 emails.

In the second iteration 100 email messages were generated and classified according to Gaussian classification and nearest neighbor method. The plot shows the variation of probability of error. It can be seen that the maximum POE is almost 0.097 in the case of nearest neighbor method and mostly the POE of the Gaussian classifier method is generally less than the nearest neighbor method. However at some instances the POE of Gaussian classification method is more at the 30th and 70th mail message (Fig. 2).

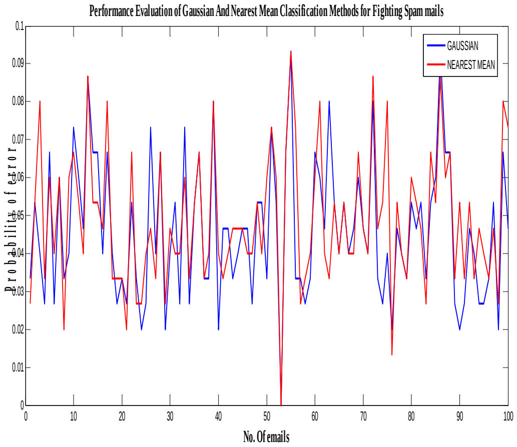

Fig. 2 It shows the variation of probability of error for 100 emails.

In the third iteration 150 email messages were generated and classified according to Gaussian classifier and nearest neighbor method. The plot shows the variation of probability of error. It can be seen that the maximum POE is almost 0.095 in the case of nearest neighbor method and mostly the POE of the Gaussian classification method is generally less than the nearest neighbor method. However at some instances the POE of Gaussian classification method is more at the 40th and 140th email message (Fig. 3).

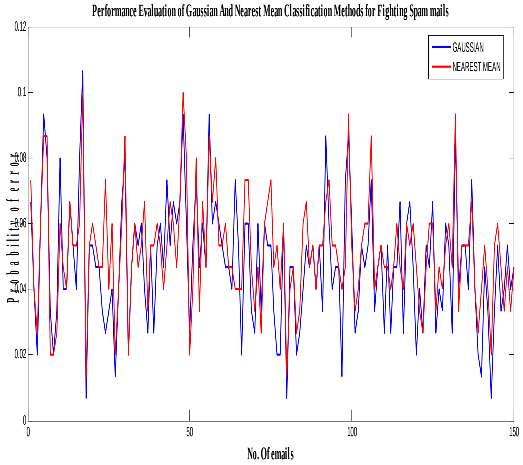

Fig. 3 It shows the variation of probability of error for 150 emails.

In the fourth iteration 200 email messages were generated and classified according to Gaussian classification and nearest neighbor method. The plot shows the variation of probability of error. It can be seen that the maximum POE is almost 0.109 in the case of nearest neighbor method and mostly the POE of the Gaussian classification method is generally less than the nearest neighbor method. However at some instances the POE of Gaussian classification method is more is at the 20th and 50th email message (Fig. 4).

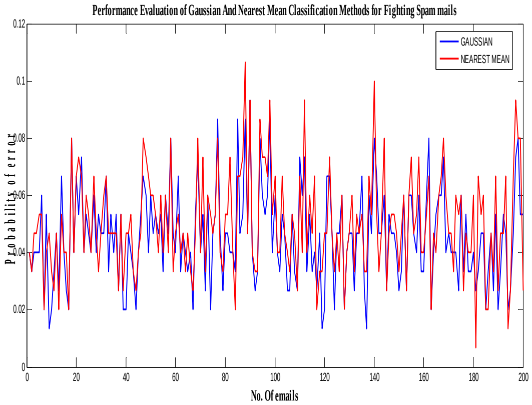

Fig. 4 It shows the variation of probability of error for 200 emails.

In the next iteration 250 email messages were generated and classified according to Gaussian classification and nearest neighbor method. The plot shows the variation of probability of error. It can be seen that the maximum POE is almost 0.11 in the case of nearest neighbor method and mostly the POE of the Gaussian classification method is generally less than the nearest neighbor method. However at some instances the POE of Gaussian classification method is more is at the 120th and 240th email message (Fig. 5).

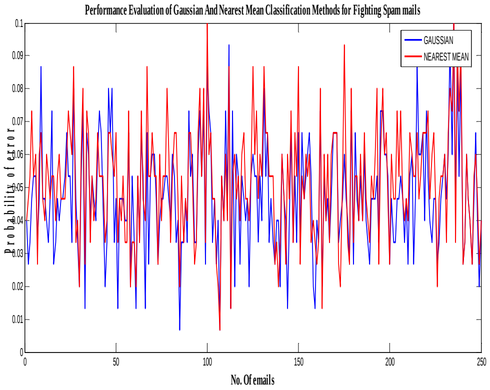

Fig. 5 It shows the variation of probability of error for 250 emails.

In the next experiment email messages were generated and classified according to Gaussian classifier and nearest neighbor method and the time taken to classify was plotted (Fig. 6).

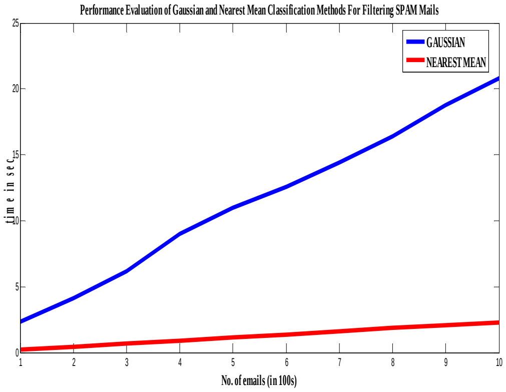

Fig. 6 It shows evaluation of Gaussian and Nearest nighbor for Filtering spam mails in time scale.

IV. Conclusion and Future Scope

It can be seen from the above iterations that most of times Gaussian classification method gives better performance and the POE is less as compared to Nearest Neighbor method. Still a few times the nearest neighbor method resulted in less POE but these instances are rare. From the traces of time taken by classifiers to classify the emails, it can be seen that at low volumes both the classifiers consume equal time but as the load of emails increases the Gaussian classifier takes more time than the nearest neighbor method. Since in spam filtering, more weightage is given to accuracy than the time taken to classify, so it can be concluded that the method of Gaussian Classification is better in classifying spam emails than the Nearest neighbor method.

Acknowledgment

The authors are grateful to Indo Global Engineering College, Abhipur, Mohali (India) for providing continuous support throughout the work.

References

[1] Zhang M, Zhang. Tjandra D, Wong STC. (2008): *DVMap: a space conscious data visualization and space discovery framework for biomedical data warehouse: Information Technology in Biomedicine*, IEEE Transactions, 8 (3), pp. 343-353.

[2] Chia HWK, Tan CL, Sung SY. (2006): *Enhancing Knowledge discovery via association based evolution of neural logic networks: Knowledge and Data Engineering*, IEEE Transactions, 18(7), pp. 889-901.

[3] Jinwook S, Shneiderman B, Sung SY. (2006): *Knowledge discovery in high dimensional data: case studies and a user survey for the rank by feature framework: Visualization and Computer Graphics*, IEEE Transactions, 12(3), pp. 311-322.

[4] Gong X, Nakamura K, Hua Y, Kei Y, Nobuhuro G. (2007) : *BAAQ: An Infrastructure for application integration and Knowledge discovery in Bioinformatics: Information Technology in Biomedicine*, IEEE Transactions, 11(4), pp. 428-434.

[5] Caulfield B., “Hallmark without the Revenue?” Internet World, 22 Feb. 1999.

[6] Evans R, Fisher D. (1994): *Overcoming process delays with Decision Tree Induction*, IEEE Expert, 9(1), pp. 60-66.

[7] Pazzani MJ. (2000): *Knowledge discovery from data. Intelligent systems and their applications*, 15(2), pp. 10-12.

[8] Longbing C, Chengqi Z. (2007): *Domain Driven actionable Discovery: Intelligent Systems*, IEEE, 22 (4), pp. 78-88.

[9] Siromoney A, Raghuram L, Korah I, Prasad GNS. (2000): *Inductive Logic programming for knowledge discovery from MRI data: Engineering in medicine and biology magazine*, IEEE, 19(4), pp. 72-77.

[10] Bojarczuk CC, Lopes HS, Frietas AA. (2000): *Genetic Programming for knowledge discovery in chest pain diagnosis: Engineering in medicine and biology magazine*, IEEE, 19(4), pp. 38-44.

[11] Tsumoto S. (2000): *Automated discovery of positive and negative knowledge in clinical database: Engineering in medicine and biology magazine*, IEEE, 19(4), pp. 56-62.

[12] Cios KJ, Pedrycz W, Swiniarsk RM. (1998): *Data mining methods for Knowledge discovery: Neural networks*, IEEE Transactions, 9(6), pp.1533-1534.

[13] Limin Fu.(1999): *Knowledge discovery by inductive neural networks: Knowledge and data engineering*, IEEE Transactions, 11(6), pp. 992-998.

[14] Pieper J, Srinivasan S, Dom B. (2001): *Streaming media knowledge discovery. Computer*, 34(9), pp. 68-74.

[15] Roddick JF, Spiliopoulou M. (2002): *A survey of temporal knowledge discovery paradigms and methods. Knowledge and data engineering*, IEEE Transactions, 14(4), pp. 750-767.

[16] Kulkarni A, McCaslin S. (2004): *Knowledge discovery from multi spectral satellite Images. Geoscience and Remote sensing letters*, IEEE, 1(4), pp. 246-250.

[17] Castro ARG, Miranda V. (2005): *Knowledge discovery in neural networks with application to transformer failure diagnosis*, IEEE Transactions, 20(2), pp. 717-724.

[18] Drucker H, Donghui W, Vapnil VN. (1999): *Support vector machines for spam categorization*, IEEE Transactions, 10(5), pp. 1048-1054.

[19] Hoanca B. (2006): “How good are our weapons in spam wars?,” Technology and Society Magazine, IEEE, 25(1), pp. 22-30.