Assessment of Performances of Spam Detection Capabilities of Nearest Mean Classifier and Gaussian Method

Srivatsan Rengamani

Department of Computer Science,
St. Joseph College of Arts and Science,
Chennai, TN, India
srivatsan81@gmail.ccom

Indian Journal of Advances in Science Engineering and Technology
Vol 1 No 1 2022: 1-7

IJASET

Review Article

Assessment of Performances of Spam Detection Capabilities of Nearest Mean Classifier and Gaussian Method

Srivatsan Rengamani
Department of Computer Science, St. Joseph College of Arts and Science, Chennai, TN, India
srivatsan81@gmail.ccom

Abstract

The potency of nearest mean classifiers and Gaussian in detecting SPAM has been investigated and the findings are presented in this paper. The outcomes take the form of probabilities of error traces and the duration of classification. Due to the difficulty of detecting SPAM, these automated techniques will save a great deal of resources and time necessary to manage messages in the inbox. Machine learning has been shown to be an effective method for classifying emails and identifying spam. It is also shown that the Gaussian Classification approach performs better at identifying messages as spam than the Nearest Mean method.

Keywords: **Nearest Mean Classifier, Knowledge Discovery, Data Mining, Gaussian Mean, Spam Detection.**

1. Introduction

The field of knowledge discovery and data mining (KDD) utilizes research results from statistical data, database systems, and knowledge based systems to develop digital instruments that empower analysts to extract knowledge from enormous databases. Individuals in science [2–4], business [1], and academic organizations as well as government bodies collect these data sets, and a number of commercial KDD packages are now available. An essential KDD objective is to "convert data into knowledge"; for instance, the knowledge gained via the application of these techniques to a digital archive could be presented in a medical journal. The knowledge gained by examining financial or commercial information could influence corporate practice or the curriculum of a university. Furthermore, information from the KDD may be used to provide justification for dismissing a credit application as required by the laws of some countries. On occasion, KDD tools are expected to assist in decision making based on gathered criteria of facts in a court of law. A case similar to that of Microsoft was sued by Mountain just for the reason that its intelligent system categorized digital greeting cards into the spam folder. In one of the earliest examples of KDD's success, Doug Fisher and Robert Evans collected information from a printing machine, identified scenarios in which the machine malfunctioned, and created guidelines to avoid these breakdowns [5, 6]. With the rising storage space profitability, the accompanying increase in the volume of data preserved, and the growing credit of historical data due to the importance of chronological data, mining sequential rules from both steady state and longitudinal data is now a possibility.

Data mining and statistical approaches to marginalize, methodical data sets, chorological mining of data, and preordered datasets are a continuation of this endeavor [7]. The process of discovering knowledge consists of commercial comprehension, detecting data necessities, acquiring and database building, methodology choosing, assessing the methodology, and implementation [8–10]. Numerous strategies are being suggested [11, 12] to enhance the runtime of algorithms that mine data. Though there are many such best performing algorithms employed in software systems of science, medicine, and commerce, science [13, 14], and medicine [15-16] have been developed, few apps have been used to combat internet spam. A few attempts have been made thus far that are costly and time-consuming experiments [17-18].

2. Materials and Methods

For this simulation experiment, MATLAB was employed as the programming language for this simulation experiment. Random samples were generated for every single category of emails, and they were divided into two equal-sized subgroups using random sampling distribution. One subgroup is considered to be a set for training while the other is considered a set for testing.

We computed the coefficients of the Gaussian probability density for each case using the training dataset of the respective class. The parameter estimations were used to calculate the Gaussian prediction equation for each case. The Gaussian classifier has been created for spam issues. The test specimens have been separated into their respective classes. For each scenario, the chance of classification error and classification time have been determined and measured. In addition, the nearest means classifier has indeed been deployed. Each class's test samples have been categorized. The time taken for classification and the classification error probability (widely known as the POE) has both been estimated for each instance. The effectiveness of the two strategies in preventing spam has also been compared, taking into account the likelihood of classification error and classification time.

3. Results and discussion

50 emails were created in the initial iteration, and they were categorized using both the methods of means of the closest and Gaussian. The variation in the probability of error or mistake has been plotted. As can be observed, the greatest POE for the nearest mean method is about 0.093, and most of the time, the probability of classification error for the Gaussian mean approach is greater than that of the nearest neighbor method. The POE of the Gaussian mean technique, however, is sometimes closer to the email messages of order 25 and 35 as shown in Figure 1.

100 emails were generated and categorized using the Gaussian mean and closest mean methods in the second iteration. The graphic displays the variation in mistake probability. As can be observed, the maximum POE for the nearest mean method is about 0.097, and the POE for the Gaussian mean approach is typically lower than that of the nearest neighbor method. However, in other cases, such as at the 30th and 70th email message, the POE of the Gaussian mean technique is greater (Figure 2).

The Gaussian mean and nearest mean approaches were used in the third iteration to create and categorize 150 emails. The image shows how the likelihood of error varies. The probability of error of the nearest neighbor method is often greater than that of the Gaussian method, and the greatest POE of the nearest mean technique is almost 0.095, which is obvious. The POE of the Gaussian mean methodology is higher in some instances, such as at email message numbers 40 and 140 as shown in Figure 3. The classification was applied once again in iteration number three, using both the methods for about 150 emails that were generated. The probability of error of the nearest neighbor method is often greater than that of the Gaussian method, and the greatest POE of the nearest mean technique is almost 0.095, which is obvious. The POE of the Gaussian mean technique is higher in some instances, such as at email message numbers 40 and 140, as shown in Figure 3.

The Gaussian mean and nearest mean approaches were used in the third iteration to create and categorize 150 emails. The image shows how the likelihood of error varies. The probability of error of the nearest neighbor method is often greater than that of the Gaussian method, and the greatest POE of the nearest mean technique is almost 0.095, which is obvious. The POE of the Gaussian mean methodology is higher in some instances, such as at email message numbers 40 and 140 as shown in Figure 3. The classification was applied once again in iteration number three, using both the methods for about 150 emails that were generated. The probability of error of the nearest neighbor method is often greater than that of the Gaussian method, and the greatest POE of the nearest mean technique is almost 0.095, which is obvious. The POE of the Gaussian mean technique is higher in some instances, such as at email message numbers 40 and 140 as shown in Figure 3.

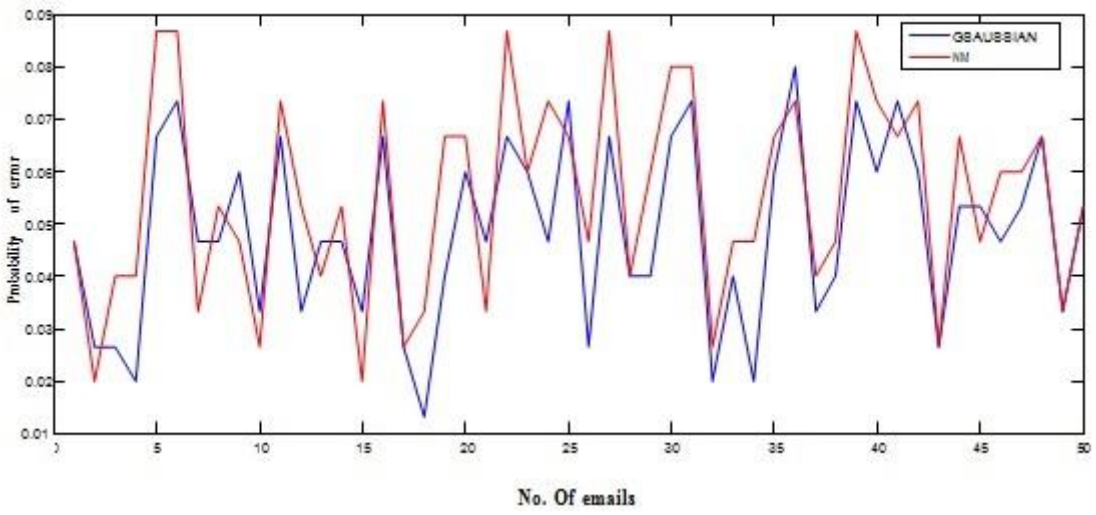

Figure 1: Probability Variation Error (50 mails)

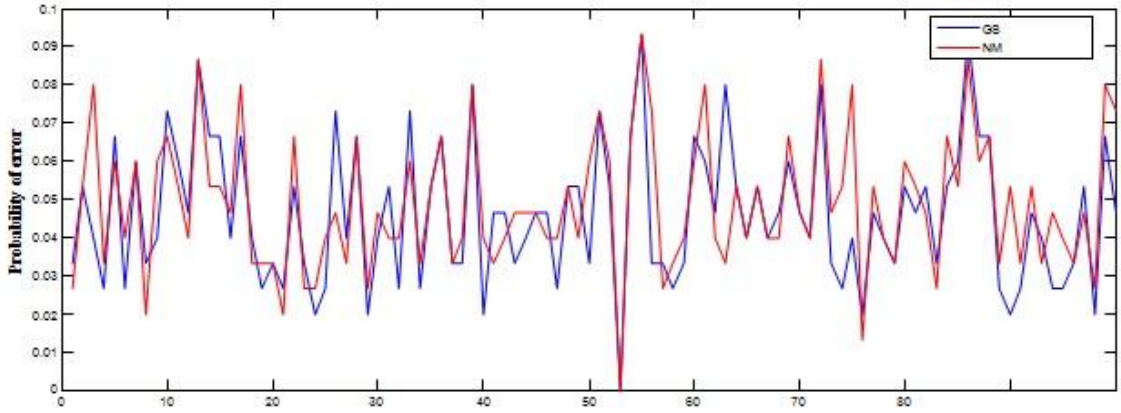

Figure 2: Probability Variation Error (100 mails)

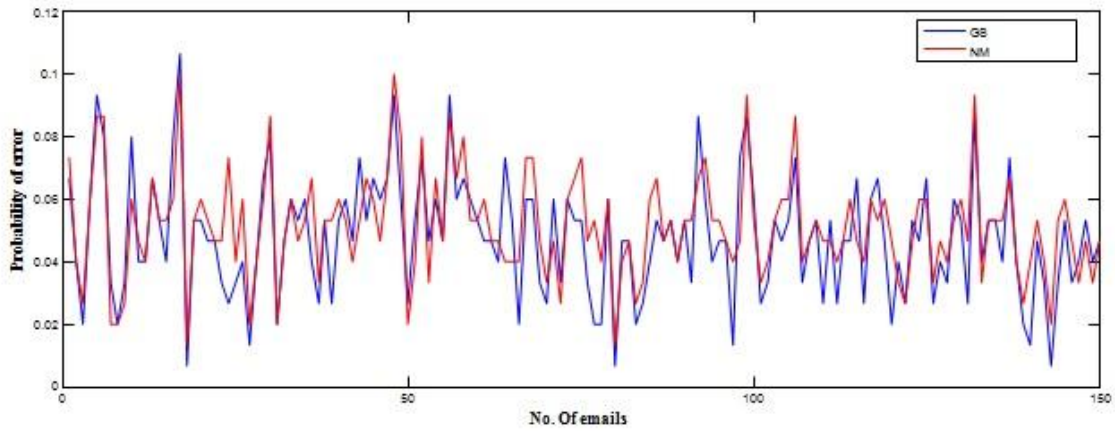

Figure 3: Probability Variation Error (150 mails)

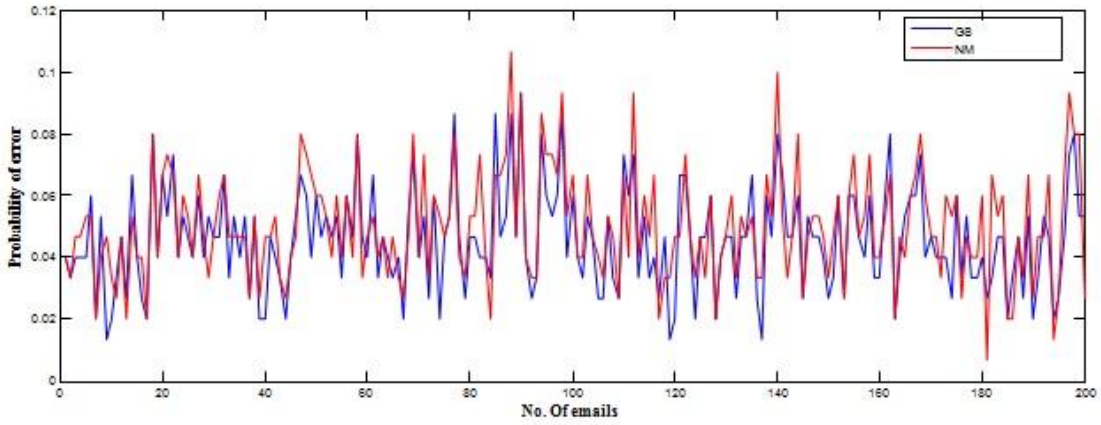

Figure 4: Probability Variation Error (200 mails)

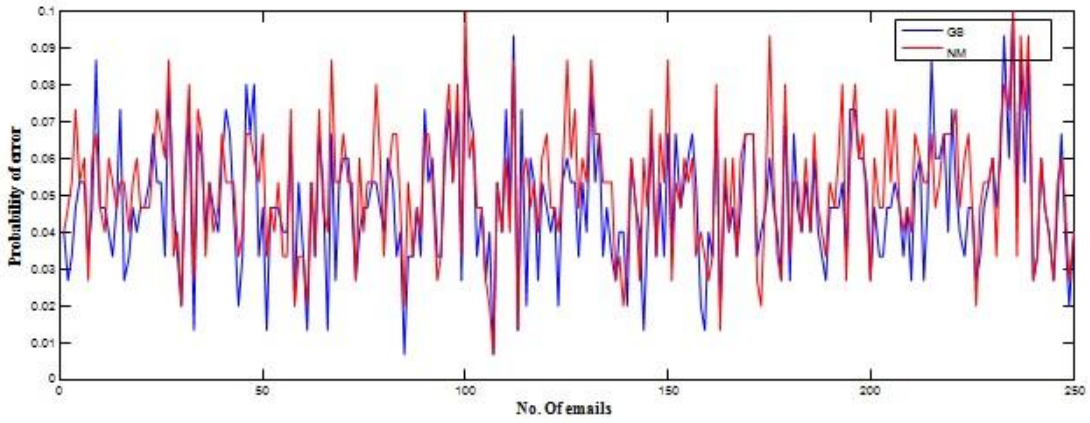

Figure 5: Probability Variation Error (250 mails)

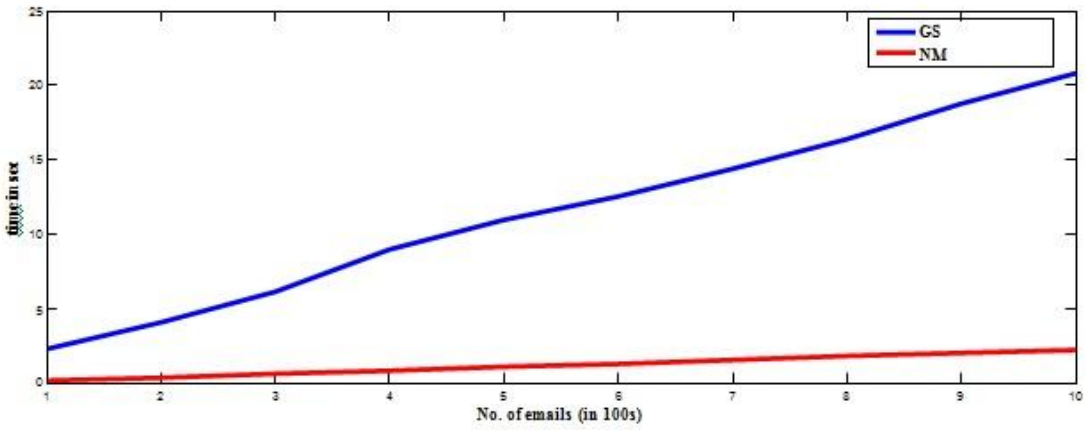

Figure 6: Time Taken for Classification

200 emails were created and categorized using the Gaussian mean and closest mean methods in the fourth iteration. The graphic displays the variation in mistake probability. As can be observed, the nearest mean technique's maximum POE is close to 0.109, and the POE of the nearest neighbor method is typically greater

than that of the Gaussian method. Similarly, in other cases, such as the 20th and 50th email message, the POE of the Gaussian mean technique is greater (Figure 4).

250 emails were created and categorized using the Gaussian mean and closest mean methods in the next iteration. The graphic displays the variation in mistake probability. As can be observed, the nearest mean technique's maximum POE is close to 0.11, while the POE of the nearest neighbor approach is typically greater than that of the Gaussian mean method. Similarly, in other cases, such as at the 120th and 240th email message, the POE of the Gaussian means technique is greater (Figure 5). Consequent experiments created and classified mail using the closest neighbor approach and the Gaussian mean, and Figure 6 shows the plotting of time taken for the classifications.

4. Conclusion

From the above discussed iterations, it can be observed that the Gaussian mean method typically performs better and has a lower POE than the Nearest Neighbor method. Even so, the nearest mean technique occasionally produces less POE, but these occurrences are uncommon. When classifying emails, classifiers take almost equal amounts of time at low volumes. However, when the quantity of messages increases, the classifier of the Gaussian approach outperforms the nearest mean technique in terms of processing speed. Since accuracy is valued more highly in spam detection than classification time, it can be deduced that the Gaussian Classification approach performs better at identifying messages as spam far better than the Nearest Mean method.

5. References

1. Zhang M, Zhang H, Tjandra D, Wong S. DBMap: A Space-Conscious Data Visualization and Knowledge Discovery Framework for Biomedical Data Warehouse. IEEE Transactions on Information Technology in Biomedicine. 2004 Sep;8(3):343–53. Available from: http://dx.doi.org/10.1109/titb.2004.832550
2. Chia H, Tan C, Sung S. Enhancing knowledge discovery via association-based evolution of neural logic networks. IEEE Transactions on Knowledge and Data Engineering. 2006 Jul;18(7):889–901. Available from: http://dx.doi.org/10.1109/tkde.2006.111
3. Jinwook Seo S, Shneiderman B. Knowledge discovery in high-dimensional data: case studies and a user survey for the rank-by-feature framework. IEEE Transactions on Visualization and Computer Graphics. 2006 May;12(3):311–22. Available from: http://dx.doi.org/10.1109/tvcg.2006.50
4. Gong X, Nakamura K, Yu H, Yura K, Go N. BAAQ: An Infrastructure for Application Integration and Knowledge Discovery in Bioinformatics. IEEE Transactions on Information Technology in Biomedicine. 2007 Jul;11(4):428–34. Available from: http://dx.doi.org/10.1109/titb.2006.888700
5. Evans B, Fisher D. Overcoming process delays with decision tree induction. IEEE Expert. 1994 Feb;9(1):60–6. Available from: http://dx.doi.org/10.1109/64.295130
6. Pazzani M. Knowledge discovery from data? IEEE Intelligent Systems. 2000 Mar;15(2):10–2. Available from: http://dx.doi.org/10.1109/5254.850821
7. Cao L, Zhang C, Yang Q, Bell D, Vlachos M, Taneri B, et al. Domain-Driven, Actionable Knowledge Discovery. IEEE Intelligent Systems. 2007 Jul;22(4):78–88, c3. Available from: http://dx.doi.org/10.1109/mis.2007.67
8. Siromoney A, Raghuram L, Siromoney A, Korah I, Prasad G. Inductive logic programming for knowledge discovery from MRI data. IEEE Engineering in Medicine and Biology Magazine. 2000 Jul;19(4):72–7. Available from: http://dx.doi.org/10.1109/51.853484
9. Bojarczuk C, Lopes H, Freitas A. Genetic programming for knowledge discovery in chest-pain diagnosis. IEEE Engineering in Medicine and Biology Magazine. 2000;19(4):38–44. Available from: http://dx.doi.org/10.1109/51.853480
10. Tsumoto S. Automated discovery of positive and negative knowledge in clinical databases. IEEE Engineering in Medicine and Biology Magazine. 2000;19(4):56–62. Available from: http://dx.doi.org/10.1109/51.853482
11. Cios K, Pedrycz W, Swiniarsk R. Data Mining Methods for Knowledge Discovery. IEEE Transactions on

Neural Networks. 1998 Nov;9(6):1533–4. Available from: http://dx.doi.org/10.1109/tnn.1998.728406
12. LiMin F. Knowledge discovery by inductive neural networks. IEEE Transactions on Knowledge and Data Engineering. 1999;11(6):992–8. Available from: http://dx.doi.org/10.1109/69.824623
13. Pieper J, Srinivasan S, Dom B. Streaming-media knowledge discovery. Computer. 2001;34(9):68–74. Available from: http://dx.doi.org/10.1109/2.947094
14. Roddick J, Spiliopoulou M. A survey of temporal knowledge discovery paradigms and methods. IEEE Transactions on Knowledge and Data Engineering. 2002 Jul;14(4):750–67. Available from: http://dx.doi.org/10.1109/tkde.2002.1019212
15. Kulkarni A, McCaslin S. Knowledge Discovery From Multispectral Satellite Images. IEEE Geoscience and Remote Sensing Letters. 2004 Oct;1(4):246–50. Available from: http://dx.doi.org/10.1109/lgrs.2004.834593
16. Castro A, Miranda V. Knowledge Discovery in Neural Networks With Application to Transformer Failure Diagnosis. IEEE Transactions on Power Systems. 2005 May;20(2):717–24. Available from: http://dx.doi.org/10.1109/tpwrs.2005.846074
17. Drucker H, Donghui W, Vapnik V. Support vector machines for spam categorization. IEEE Transactions on Neural Networks. 1999;10(5):1048–54. Available from: http://dx.doi.org/10.1109/72.788645
18. Hoanca B. How good are our weapons in the spam wars? IEEE Technology and Society Magazine. 2006;25(1):22–30. Available from: http://dx.doi.org/10.1109/mtas.2006.1607720