Support Vector Machines Parameter Selection Based on Combined Taguchi Method and Staelin Method for E-mail Spam Filtering

Wei-Chih Hsu [1], Tsan-Ying Yu [2,*]

[1] Department of Computer and Communication, National Kaohsiung First University of Science and Technology, Kaohsiung City, Taiwan, ROC

[2] Department of Electrical Engineering, Kao Yuan University, Kaohsiung, Taiwan, ROC.

Received 13 December 2011; received in revised form 20 January 2012; accepted 09 February 2012

Abstract

Support vector machines (SVM) are a powerful tool for building good spam filtering models. However, the performance of the model depends on parameter selection. Parameter selection of SVM will affect classification performance seriously during training process. In this study, we use combined Taguchi method and Staelin method to optimize the SVM-based E-mail Spam Filtering model and promote spam filtering accuracy. We compare it with other parameters optimization methods, such as grid search. Six real-world mail data sets are selected to demonstrate the effectiveness and feasibility of the method. The results show that our proposed methods can find the effective model with high classification accuracy

Keywords: Support Vector Machines, Taguchi Method, Grid Search

References

[1] W. W. Cohen, "Fast effective rule induction," in *Proceedings of the Twelfth International Conference on Machine Learning*, 1995, pp. 115-123.

[2] W. W. Cohen, "Learning rules that classify e-mail," in *Proceedings of the 1996 AAAI Spring Symposium in Information Access*, 1996, pp. 18-25.

[3] I. Androutsopoulos, G. Paliouras, and E. Michelakis, "Learning to filter unsolicited commercial e-mail," " DEMOKRITOS", National Center for Scientific Research Technical report 2004/2, 2004.

[4] M. Collins, R. E. Schapire, Y. Singer, P. Domingos, W. Fan, S. J. Stolfo, J. Zhang, P. K. Chan, Y. Freund, and R. Schapire, "Boosting Trees for Anti-Spam Email Filtering," *4th International Conference on Recent Advances in Natural Language Processing*, 2001, pp. 1189-1232.

[5] I. Androutsopoulos, G. Paliouras, V. Karkaletsis, G. Sakkis, C. D. Spyropoulos, and P. Stamatopoulos, "Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach," presented at the Proceedings of the workshop "Machine Learning and Textual Information Access", 4th European Conference on Principles and Practice of Knowledge Discovery in Databases, 2000.

[6] V. N. Vapnik, *The nature of statistical learning theory*. New York: Springer Verlag, 2000.

[7] J. Provost, "Naive-bayes vs. rule-learning in classification of email. The University of Texas at Austin," Artificial Intelligence Lab. Technical Report AI-TR-99-284, 1999.

[8] C. L. Huang and C. J. Wang, "A GA-based feature selection and parameters optimizationfor support vector machines," *Expert Systems With Applications*, vol. 31, pp. 231-240, 2006.

* Corresponding author. E-mail address: allen@nfu.edu.tw
Tel.: +886-5-6315368; Fax: +886-5-6314486

[9] T. Howley and M. G. Madden, "The genetic kernel support vector machine: Description and evaluation," *Artificial Intelligence Review*, vol. 24, pp. 379-395, 2005.
[10] G. Taguchi and S. Chowdhury, *Robust engineering*, New Work: McGraw-Hill, 2000.
[11] C. C. Chang and C. J. Lin. (2008). *LIBSVM -- A Library for Support Vector Machines*. http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[12] G. Taguchi, *Introduction to quality engineering*, Tokyo: Asian Productivity Organization, 1990.
[13] M. Phadke, *Quality engineering using robust design*, U.S.A: Prentice Hall PTR Upper Saddle River, 1995.
[14] D. C. Montgomery, *Design and analysis of experiments*, New York: Wiley, 2006.
[15] C. Staelin, "Parameter selection for support vector machines," *Hewlett-Packard Company, Tech. Rep. HPL-2002-354R1*, 2003.
[16] N. Logothetis and H. P. Wynn, *Quality through design: experimental design, off-line quality control, and Taguchi's contributions*, Oxford: Clarendon Press, 1989.
[17] *Enronspam*. http://www.aueb.gr/Users/ion/data/enron-spam/