

# Expediting Model Selection for Support Vector Machines Based on Data Reduction

Yu-Yen Ou, Chien-Yu Chen, Shien-Ching Hwang and Yen-Jen Oyang

Department of Computer Science and Information Engineering

National Taiwan University, Taipei, Taiwan

{yien,yjoyang}@csie.ntu.edu.tw

{cychen,schwang}@mars.csie.ntu.edu.tw

The support vector machine was first proposed by Vapnik [1] and has since attracted a high degree of interest in the machine learning research community. Several recent studies have reported that the SVM (support vector machines) generally are capable of delivering higher performance in terms of classification accuracy than the other data classification algorithms. However, for some datasets, the performance of SVM is very sensitive to how the cost parameter and kernel parameters are set. As a result, the user normally needs to conduct extensive cross validation in order to figure out the optimal parameter setting. This process is commonly referred to as model selection.

One practical issue with model selection is that this process is very time-consuming. For example, if the model selection procedure adopted in [2] is employed to construct a SVM for the `shuttle` dataset in the `Statlog` collection [3], then the complete grid search model selection process will take over 60 hours on a machine equipped with dual pentium-1GHz CPUs and 1GB RAM. As the `shuttle` dataset, containing 43500 training samples, is not considered as a large case in the contemporary environment, how to speed up the model selection process for SVM becomes a crucial issue and several studies have been conducted to address this issue in recent years [4], [5], [6], [7]. These studies share a common ground aimed at reducing the search space of parameter combinations.

In this paper, a data reduction based approach aimed at expediting the model selection process of SVM is proposed. The main idea of the proposed approach is to employ a data reduction mechanism to remove the non-essential training samples and thus reduce the size of the training dataset. Experimental results reveal that the data reduction based approach is able to speed up the model selection process by around 2 times to 4000 times, depending on the characteristics of the datasets. Furthermore, the experimental results reveal that the classification accuracy of SVM is not traded by employing the proposed approach. In other words, the SVM with the parameter setting determined by the proposed approach is able to achieve the same level of classification accuracy as the SVM with the parameter setting determined by the traditional model selection process. In these experiments, the LIBSVM [8], an integrated software for support vector machines, with the RBF kernel is employed. The datasets employed include `satimage`, `letter`, and `shuttle` from the `Statlog` collection and `isolet` from the UCI Repository [9].

As far as the execution time of the proposed approach is concerned, the average time complexity of the data reduction process is  $O(n \log n)$ , where  $n$  is the number of samples in the training dataset. In the experiments reported in this paper, the execution times of the data reduction process are 37.6 seconds, 259.4 seconds, 711.7 seconds, and 1412.9 seconds, for `satimage`, `letter`, `shuttle`, and `isolet`, respectively. If we add the time taken to carry out data reduction and the time taken to carry out model selection with the reduced dataset, then the total time taken to determine the parameter setting for the `shuttle` dataset is  $711.7 + 254.7 = 966.4$  seconds, versus over 60 hours taken to carry out model selection with the original dataset.

## REFERENCES

- [1] C. Cortes and V. Vapnik, "Support-vector network," *Machine Learning*, vol. 20, pp. 273–297, 1995.
- [2] K.-M. Chung, W.-C. Kao, Tony Sun, and C.-J. Lin, "Decomposition methods for linear support vector machines," Tech. Rep., Department of Computer Science and Information Engineering, National Taiwan University, 2002.
- [3] D. Michie, D. J. Spiegelhalter, and C. C. Taylor, *Machine Learning, Neural and Statistical Classification*, Prentice Hall, Englewood Cliffs, N.J., 1994, Data available at <ftp://ftp.ncc.up.pt/pub/statlog/>.
- [4] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Machine Learning*, vol. 46, pp. 131–159, 2002.
- [5] S. Sathya Keerthi, "Efficient tuning of SVM hyperparameters using radius/margin bound and iterative algorithms," *IEEE Transactions on Neural Networks*, 2002, To appear.
- [6] K. Duan, S. S. Keerthi, and A. N. Poo, "Evaluation of simple performance measures for tuning SVM hyperparameters," *Neurocomputing*, 2002, To appear.
- [7] D. DeCoste and K. Wagstaff, "Alpha seeding for support vector machines," in *Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD-2000)*, 2000.
- [8] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

- [9] C. L. Blake and C. J. Merz, "UCI repository of machine learning databases," Tech. Rep., University of California, Department of Information and Computer Science, Irvine, CA, 1998, Available at <http://www.ics.uci.edu/~mllearn/MLRepository.html>.