Parallel Cost-Sensitive Support Vector Machine Software for Classification

Tatjana Eitrich, Bruno Lang

published in

NIC Workshop 2006,
From Computational Biophysics to Systems Biology,
Jan Meinke, Olav Zimmermann,
Sandipan Mohanty, Ulrich H.E. Hansmann (Editors)
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **34**, ISBN-10: 3-9810843-0-6,
ISBN-13: 978-3-9810843-0-6, pp. 141-144, 2006.

http://www.fz-juelich.de/nic-series/volume34

Parallel Cost-Sensitive Support Vector Machine Software for Classification

Tatjana Eitrich[1] **and Bruno Lang**[2]

[1] Central Institute for Applied Mathematics,
Research Centre Jülich, 52425 Jülich, Germany
E-mail: t.eitrich@fz-juelich.de

[2] Applied Computer Science and Scientific Computing Group,
Department of Mathematics, University of Wuppertal, Germany
E-mail: lang@math.uni-wuppertal.de

1 Support Vector Machines

Support vector machines (SVMs) are well-known data mining methods for classification and regression problems[1]. Their popularity is mainly due to their applicability in various fields of data mining, such as text mining[2], biomedical research[3], and many more. Their accuracy is excellent and in many cases they outperform other machine learning methods such as neural networks. SVMs have their roots in the field of statistical learning which provides the reliable generalization theory[4]. Several properties that make this learning method successful are well-known, e.g. the kernel trick[5] for nonlinear classification and the sparse structure of the final classification function. In addition, SVMs have an intuitive geometrical interpretation, and a global minimum can be located during the SVM training phase. In comparison to genetic algorithms or neural networks, less experience is required for using them, which helps researchers to get started with SVM software quite fast.

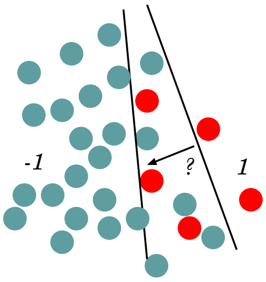

Figure 1. The problem of unbalanced and cost-sensitive classification.

2 Cost-Sensitive Support Vector Machine

Data sets with different class distributions lead to the effect that conventional machine learning methods are biased towards the larger class[6,7]. To overcome this problem and to obtain sensitive but also accurate classifiers we extended and improved the standard SVM formulation. In addition we use techniques addressing the problem of unbalanced classification, such as oversampling and threshold moving[8]. In Fig. 1 an unbalanced toy problem is shown. Recent results for a CYP P450 drug classification problem are given in ref 9.

Parallel Support Vector Machine

The main drawback of current SVM models is their high computational complexity for large data sets[10]. This can in fact restrict the applicability of SVMs since the amount of data for classification modeling increases dramatically. Therefore the development of highly scalable parallel SVM algorithms is a new important topic of current SVM research. Some algorithms for parallel SVM learning already do exist, but most of them are limited to heuristics for distributed training on reduced data sets[11, 12]. These are not useful as stand-alone systems for high quality learning on large data. We have implemented a parallel support vector machine software well suited for multi-processor shared memory (SMP) clusters that become more and more available. Our algorithm can be used in serial and parallel mode. The parallel implementation provides pure MPI and OpenMP modes as well as a hybrid mode which combines fine and coarse grained parallelization aspects to a well scalable SVM learning method[13]. The fine grained inner parallel scheme is shown in Fig. 2.

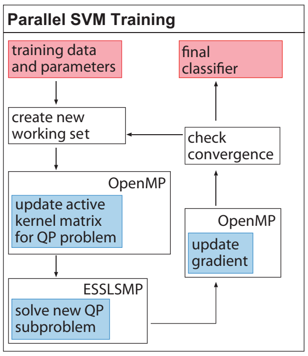

Figure 2. Parallel training algorithm based on a decomposition method for SVM training.

Summary and Future Work

We obtained a flexible cost-sensitive parallel SVM software that can be used on high-end machines with SMP architectures to process the large data sets that arise more and more in bioinformatics and other fields of research. Future work will be on enhanced parameter tuning.

Acknowledgments

The computations were performed on the IBM p690 cluster JUMP at Research Centre Jülich. The authors would like to thank the ZAM team at Jülich for support.

References

1. N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, Cambridge, UK, 2000.
2. T. Joachims. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning (ECML'98), Chemnitz*, volume 1398 of *Lecture Notes in Computer Science*, pages 137–142. Springer, 1998.
3. H. Yu, J. Yang, W. Wang, and J. Han. Discovering compact and highly discriminative features or feature combinations of drug activities using support vector machines. In *2nd IEEE Computer Society Bioinformatics Conference (CSB 2003), 11-14 August 2003, Stanford, CA, USA*, pages 220–228. IEEE Computer Society, 2003.
4. V. N. Vapnik. *Statistical learning theory*. John Wiley & Sons, New York, 1998.
5. B. Schölkopf. The kernel trick for distances. In *NIPS*, pages 301–307, 2000.
6. C. Elkan. The foundations of cost-sensitive learning. In *Proc. of the 17. International Joint Conference on Artificial Intelligence (IJCAI'01)*, pages 973–978, 2001.
7. M. A. Maloof. Learning when data sets are imbalanced and when costs are unequal and unknown. In *Worksop on Learning from Imbalanced Data Sets II, ICML, Washington DC*, 2003.
8. R. Akbani, S. Kwek, and N. Japkowicz. Applying support vector machines to imbalanced datasets. In *Proceedings of the 15th European Conference on Machine Learning (ECML'04), Pisa, Italy, 2004*, volume 3201 of *Lecture Notes in Computer Science*, pages 39–50. Springer, 2004.
9. T. Eitrich, A. Kless, C. Druska, J. Grotendorst, and W. Meyer. Classification of highly unbalanced CYP450 data of drugs using cost sensitive machine learning techniques. Technical Report FZJ-ZAM-IB-2006-09, FZJ, 2006.
10. N. Chen, W. Lu, J. Yang, and G. Li. *Support vector machine in chemistry*. World Scientific Pub Co Inc, 2004.
11. H. P. Graf, E. Cosatto, L. Bottou, I. Dourdanovic, and V. Vapnik. Parallel support vector machines: the cascade SVM. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 521–528. MIT Press, Cambridge, MA, 2005.
12. J.-X. Dong, A. Krzyzak, and C. Y. Suen. A fast parallel optimization for training support vector machines. In P. Perner and A. Rosenfeld, editors, *Proceedings of 3rd International Conference on Machine Learning and Data Mining*, pages 96–105, 2003.
13. T. Eitrich, W. Frings, and B. Lang. HyParSVM – a new hybrid parallel software for support vector machine learning on SMP clusters. In *Proceedings of the Euro-Par Conference, Dresden*, 2006.