Nonparametric Classification with Polynomial MPMC Cascades[1]

Sander M. Bohte [a] Markus Breitenbach [b]

Gregory Z. Grudic [b]

[a] CWI, Kruislaan 413, 1098 SJ Amsterdam
[b] University of Colorado, Boulder, Colorado, USA

This paper proposes a computationally efficient class of nonparametric binary classification algorithms that generate nonlinear separating boundaries, with minimal tuning of learning parameters. We avoid the computational pitfalls of using extensive cross validation for model selection. For example, in Support Vector Machines (SVMs) [6], both the choice of kernels and corresponding kernel parameters is based on extensive cross validation experiments, making generating good SVM models computationally very difficult. Other algorithms, such as Minimax Probability Machine Classification (MPMC) [5], Neural Networks, and even ensemble methods such as Boosting, can suffer from the same computational pitfalls.

The Minimax Probability Machine for Classification (MPMC), due to Lanckriet *et al.* [5], is a recent algorithm that has this characteristic. Given the means and covariance matrices of two classes, MPMC calculates a hyperplane that separates the data by minimizing the maximum probability of misclassification. As such, it generates both a classification and a bound on the expected error for future data. In the same paper, the MPMC is also extended to non-linear separating hypersurfaces using kernel methods. However, MPMC then has similar complexity as SVM algorithms.

We propose an efficient, scalable, nonparametric approach to generating nonlinear classifiers based on the MPMC framework: the class of Polynomial MPMC Cascades (PMCs). PMCs are motivated by cascading algorithms like cascade-correlation [1] and Tower [3], which sequentially add levels that improve performance. However these algorithms applied to real world problems often suffer from overfitting and generally scale poorly to larger problems due to the increasing number of variables used in subsequent levels of the cascades.

In our cascading algorithm, for levels, instead of neural networks, we use low dimensional polynomials, after Grudic & Lawrence's Polynomial Cascade algorithm for regression [4], which efficiently builds very high dimensional nonlinear regression surfaces using cascades of such polynomials. In this manner, we avoid the growth in learning complexity in cascade-correlation type algorithms by always projecting the intermediate outputs onto a two-dimensional state-space, and proceed from there. Since the rate of convergence (as a function of the number of training examples) of a basis function learning algorithm depends on the size of the

[1] Full paper to appear in the proceedings of ICML 2004, July 4-8 2004, Banff, Canada

state-space (i.e. the number of basis functions), these low dimensional projections lead to stable cascade models.

The proposed Polynomial MPMC Cascade algorithms generate a nonlinear hypersurface from a cascade of low-dimensional polynomial structures. The optimal choice for each level of the cascade is determined using MPMC to select the next most discriminating structure. From one level to the next, these additional discriminating structures are added to the cascade using MPMC, such that at each step we obtain the next most discriminating polynomial cascade; we construct PMC variants that use different ways of constructing the initial polynomial structures. By using MPMC to guide the addition of new cascade levels, we maintain a current performance *and* current maximum error bound during construction. The addition of new structures to the cascade is stopped when the error bound no longer improves.

We show that the proposed PMC algorithms yield competitive results on benchmark problems, while providing maximum error bounds. The PMCs are efficient in that their complexity is 1) linear in the number of input-dimensions, 2) linear in the number of training examples, 3) linear in the number of levels of the cascade. Additionally, the PMC algorithm is efficient in that there are no parameters that need fine-tuning for optimal performance. Unlike approaches like boosting [2], the PMC generates a single model, a polynomial, instead of an ensemble of several models while still being fast and scalable to large datasets. We recently completed work on classification of datasets with many millions of datapoints, demonstrating excellent classification performance while being linear in computing time with respect to the number of datapoints [Breitenbach, Bohte & Grudic, submitted].

To summarize, we believe that the contribution of this paper lies in effectiveness and speed of the proposed class of PMC algorithms: while being solidly rooted in the theory of MPMC, their linear time complexity and nonparametric nature allow them to essentially be a "plug & play" solution for classification problems, yielding results competitive with algorithms like MPMC with Gaussian kernels and non-linear SVMs.

References

[1] S.E. Fahlman and C. Lebiere. The cascade-correlation learning architecture. In D. S. Touretzky, editor, *NIPS*, volume 2, pages 524–532, Denver 1989.

[2] Y Freund. An adaptive version of the boost by majority algorithm. In *COLT: Proc. Workshop on Comp. Learning Theory*, 1999.

[3] S. Gallant. Perceptron-based learning algorithms. *IEEE Trans. on Neural Networks*, 1(2):179–191, 1990.

[4] G.Z. Grudic and P.D. Lawrence. Is nonparametric learning practical in very high dimensional spaces? In *Proc. 15th Intern. Joint Conf. on AI (IJCAI-97)*, pages 804–809, San Francisco, 1997.

[5] G. Lanckriet, L. El Ghaoui, C. Bhattacharyya, and M.I. Jordan. A robust minimax approach to classification. *J. of Machine Learning Research*, 3:555–582, Dec 2002.

[6] B. Schölkopf and A. Smola. *Learning with Kernels.* MIT Press, MA, 2002.