Combining Kernel Machines Through Decorrelation

Luis Pérez-Breva

Artificial Intelligence Laboratory and
The Center for Biological and Computational Learning
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: Develop an algorithm to combine kernel machines within the "decorrelating framework" [6]. Then prove generalization ability of the ensemble.

Motivation: Combinations of classifiers have been found useful empirically, yet no formal proof exists about their generalization ability. Our goal is to develop a combination of kernel machines for which it is possible to prove generalization bounds. We believe that this is possible by further elaborating the arguments presented in [6], which may provide insights on boosting methods and view-based classifiers.

Previous Work: Boosting [7], is a simple scheme to combine 3 classifiers (weak learners) by majority vote with the goal of possibly outperforming a single one. [3, 9] provide insights that relate boosting to game theory and margin classifiers.
[5, 6], present a new scheme to combine 3 classifiers by majority vote, which replaces distribution reweighting from boosting with explicit decorrelation of two classifiers. [6] proofs equivalence between this setting and boosting and suggests a connection to AdaBoost (the first implementation of boosting [2, 8]).

Neither *Boosting* nor *decorrelating classifiers* schemes provide theoretical arguments that explain their ability to generalize.

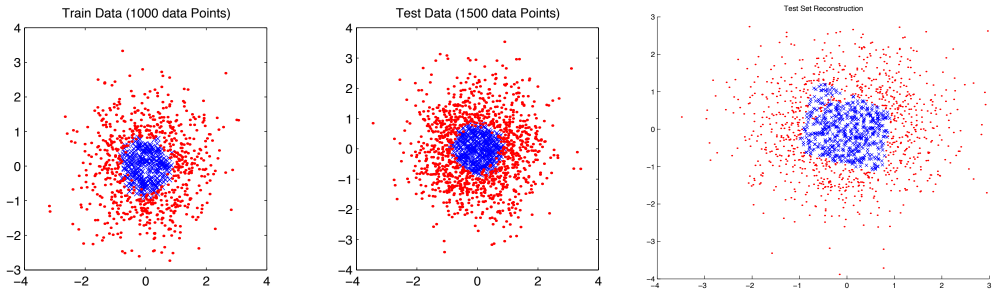
Approach: We investigate the problem of decorrelating classifiers from the set theory perspective, and prove that weak learnability is not a necessary requirement. We start our study with a preliminary approach that uses Support Vector Classifiers and implements the decorrelating classifiers scheme through a threshold. Then we present an algorithm to train a cascade of kernel machines and will take advantage of the results from [1] to prove its generalization ability.

Impact: On the practical side, our algorithm has a wide range of possible applications: simplifying kernel selection, view-based classification methods. From the theoretical perspective, a boosting-type algorithm whose ability to generalize is proven is obviously interesting.

Preliminary Results: We have carried on preliminary experiments using an early version of an algorithm to combine Linear Support Vector Machines. Figure 1 shows synthetic data we analyzed, and the prediction of the algorithm on synthetic test data. The cascade of linear SVM classifiers as currently defined is already able to outperform a single Linear SVM in both train and testing, and approximates non-linear decision surfaces by combining linear classifiers. When ran on a Face Dataset ([4]), the combined scheme still outperformed a single linear classifier (see figure 2), although failed to achieve same performance than single higher degree polynomials on the same problem.

Future Work: Provide a formal representation of the problem the algorithm is currently trying to solve. Improve the current implementation and benchmark its performance on synthetic and real data. Analyze the method to prove generalization bounds.

Research Support: Research at CBCL is sponsored by grants from: Office of Naval Research (DARPA) under contract No. N00014-00-1-0907, National Science Foundation (ITR) under contract No. IIS-0085836, National Science Foundation (KDI) under contract No. DMS-9872936, and National Science Foundation under contract No. IIS-9800032 Additional support was provided by: Central Research Institute of Electric Power Industry, Center for e-Business (MIT), Eastman Kodak Company, DaimlerChrysler AG, Compaq, Honda R&D Co., Ltd., Komatsu Ltd., Merrill-Lynch, NEC Fund, Nippon Telegraph & Telephone, Siemens Corporate Research, Inc., Toyota Motor Corporation and The Whitaker Foundation.



(a)

(b)

Figure 1: **(a)** Is obtained by random sampling from a $0-$mean $2-$dimensional gaussian distribution with $\sigma = 1$. Within a $0.8 \pm \epsilon$ from the center of the distribution class label is $+1$, and -1 elsewhere. ϵ is random $0 - mean$ white noise with $\sigma = 0.1$. The problem is clearly non linearly separable. **(b)** shows the reconstruction of test data using the preliminary version of the algorithm. Note that the algorithm is able to approximate non-linear decision surfaces using only linear classifiers.

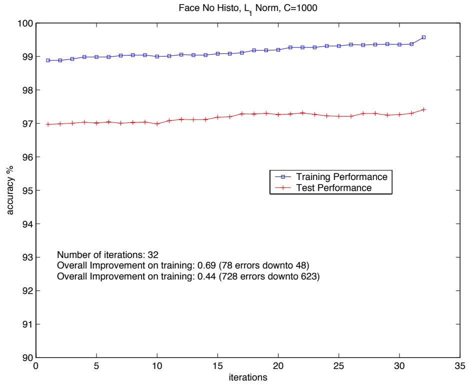


Figure 2: Results on the face dataset(361-dimensions). An SVM using 2nd degree polynomial kernel did only 520 errors on the same test set.

References:

[1] T. Evgeniou, M. Pontil, and A. Elisseeff. Leave one out error, stability, and generalization of voting combinations of kernel classifiers. Technical report, INSEAD, 2001.

[2] Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121(2):256–285, September 1995.

[3] Y. Freund and R.E. Schapire. Game theory, on-line prediction and boosting. In *Proc. 9th Annu. Conf. on Comput. Learning Theory*, pages 325–332. ACM Press, New York, NY, 1996.

[4] B. Heisele, T. Poggio, and M. Pontil. Face detection in still gray images. A.I. memo 1687, Center for Biological and Computational Learning, MIT, Cambridge, MA, 2000.

[5] P. Niyogi, J. B. Pierrot, and O. Siohan. Multiple classifiers by constrained minimization. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2000.

[6] P. Niyogi, J. B. Pierrot, and O. Siohan. On decorrelating classifiers and combining them. Unpublished, 2001.

[7] R.E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

[8] R.E. Schapire. A brief introduction to boosting. In *Proc. 16th International Joint Conference on Artificial Intelligence*, 1999.

[9] R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. In *Proc. 14th International Conference on Machine Learning*, pages 322–330. Morgan Kaufmann, 1997.