Combining Classifiers

Alexander Rakhlin

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: Combining different classifiers in a hieriarchical manner. In particular, investigating how additional information in the training set can be used to build a hierarchy of classifiers. Looking at the geometry of the problem in the feature space. Examining the tradeoff between having a single classifier trained on a set of points and a collection of classifiers trained on subsets of these points (keeping the complexity of the classifiers constant).

Motivation: In the face-recognition or face-detection domain, for instance, much more "higher-level" information about the training set is available. For example, we can label faces as 'frontal' or 'profiles'. With appropriate feature space, it is feasible that these two distributions differ enough and so two separate classifiers can be trained and then combined. In general, hierarchical organization of classifiers is an attractive approach because it is biologically-inspired (i.e. increasing specificity of neurons in the visual cortex).

Previous Work: A number of approaches have been developed in the past decade to address the question of combining classifiers. These approaches include bagging introduced by Breiman [2], boosting (Schapire [4]), decorrelating techniques (Niyogi [3]), etc. While these approaches have interesting properties, they do not take into account additional information from the training set. Our approach to the problem from a geometric point of view is inspired by the work of Vapnik and Bottou on local learning algorithms [6, 1] as well as by the work of Schneiderman et al [5] in which detectors specialized to particular views are combined.

Approach: We look at different ways to combine classifiers trained on separate datasets. For a test sample, a combination of classifiers can be used to determine the label. Alternatively, probability of membership to a particular class can be estimated and then a classifier for that class can be used to determine the label. For example, two classifiers are trained: one on the frontal faces (against non-faces) and the second on profiles. If the distributions of the two classes (frontal faces and profiles) are different enough, we might try to determine the type of the face by calculating a distance to the means or medians of two clusters and then using the trained classifier for that class. This approach reduces the complexity of the single classifiers, but suffers from having a reduced number of training points. This trade-off is important as we are working with sparse data in a high-dimensional space.

Impact: The immediate impact of the work on combining classifiers is building systems with better predictive power. The long-term goal of the project is to approach the problem of building a hierarchy of classifiers, which might be inspired by the organization in the brain and the visual cortex in particular.

Future Work: Doing more experiments with real data as well as setting the problem in a more formal theoretical framework to get VC-type or other bounds on the predictive power.

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.

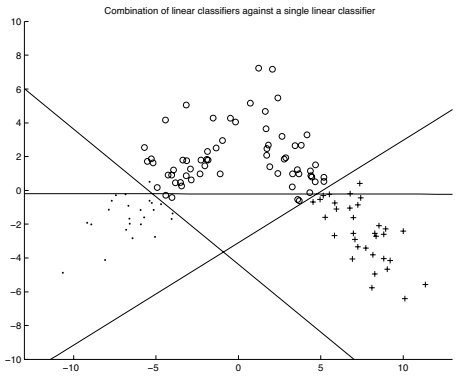

Figure 1: Toy example with two distributions

References:

[1] L. Bottou and V.N. Vapnik. Local learning algorithms. *Neural Computation*, 4(6):888–900, 1992.

[2] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[3] Partha Niyogi, Jean-Benoit Pierrot, and Olivier Siohan. On decorrelating classifiers and combining them.

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

[5] H. Schneiderman. A statistical approach to 3d object detection applied to faces and cars, 2000.

[6] V. Vapnik and L. Bottou. Local algorithms for pattern recognition and dependencies estimation. *Neural Computation*, 5:893–909, 1993.