The Sources of Increased Accuracy for Two Proposed Boosting Algorithms

David B. Skalak
Dept. of Computer Science
University of Massachusetts
Amherst, MA 01003-4610
skalak@cs.umass.edu

Abstract

We introduce two boosting algorithms that aim to increase the generalization accuracy of a given classifier by incorporating it as a level-0 component in a stacked generalizer. Both algorithms construct a complementary level-0 classifier that can only generate coarse hypotheses for the training data. We show that the two algorithms boost generalization accuracy on a representative collection of data sets. The two algorithms are distinguished in that one of them modifies the class targets of selected training instances in order to train the complementary classifier. We show that the two algorithms achieve approximately equal generalization accuracy, but that they create complementary classifiers that display different degrees of accuracy and diversity. Our study provides evidence that it may be useful to investigate families of boosting algorithms that incorporate varying levels of accuracy and diversity, so as to achieve an appropriate mix for a given task and domain.

Introduction

It is now well-established that combining classifiers can lead to increased generalization accuracy for a variety of classifier model classes and domains (*e.g.*, (Selfridge 1959; Wolpert 1992; Chan and Stolfo 1995; Breiman 1992; Breiman 1994a; Kong and Dietterich 1995; Ali and Pazzani 1995; Zhang *et al.* 1992; Jacobs *et al.* 1991; Maclin and Shavlik 1995; Skalak 1995)). In particular, the accuracy of some classifiers can be boosted by combining their predictions with those of other classifiers (Schapire 1990; Drucker *et al.* 1994; Drucker and Cortes 1996; Freund and Schapire 1995). A logical next research task is to identify the sources of increased accuracy for the various techniques that have been developed.

We first present two algorithms for boosting the accuracy of a nearest neighbor classifier (the *base classifier*) by incorporating it as a level-0 component in a stacked generalizer. The two algorithms create a nearest neighbor *complementary classifier*, which is used as an additional level-0 component. We then show that the sources of their power appear to be different, although they demonstrate comparable generalization accuracy, and although they are members of a family of related algorithms. We find that the two algorithms effect different combinations of the two factors that have been recognized as affecting the performance of a stacked generalizer: component accuracy and component diversity (Hansen and Salamon 1990; Wolpert 1992; Breiman 1992; Battiti and Colla 1994; Chan and Stolfo 1995; Ali and Pazzani 1995; Opitz and Shavlik 1995; Krogh and Vedelsby 1995). This analysis suggests that it may be useful to investigate families of boosting algorithms to facilitate tailoring the combination of those two factors to the application domain.

Algorithms

The two algorithms we introduce, *Coarse Reclassification* and *Deliberate Misclassification*, filter and/or manipulate a training set to create a complementary level-0 classifier. In this respect, this work partly falls in the footsteps of the progress shown by Chan and Stolfo (Chan and Stolfo 1993a; Chan and Stolfo 1993b; Chan and Stolfo 1995) in creating meta-level training sets for composite classifiers. However, the algorithms we present to select training instances are different from the meta-level training strategies investigated by Chan and Stolfo. The Coarse Reclassification algorithm takes training data filtering to an (efficient) extreme by using only one instance per class to support the classifier. The heuristic Deliberate Misclassification algorithm modifies the class target labels of selected training instances. As far as our review of the literature has shown, these approaches have not been investigated by any previous research.

The focus of the two algorithms we investigate for creating stacked generalizers is the creation of a complementary level-0 classifier, as shown in Figure 1. In this implementation, the only data that are passed from level-0 to level-1 are the class label predictions of the level-0 classifiers.

Both algorithms create a complementary level-0 classifier drawn from a model class of coarse-hypothesis or *high-bias* (Breiman 1994b; Kong and Dietterich

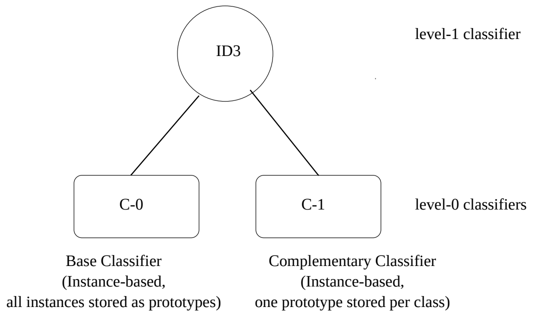

Figure 1: Stacked generalization architecture for this paper.

1995) nearest neighbor classifiers that incorporate only one prototype for each class. These algorithms are described in detail in a longer piece (Skalak 1996). Both algorithms train a complementary nearest neighbor classifier by taking a small number of random samples of prototype sets (*e.g.,* 100), each set containing one prototype from each class, and retaining the prototypes that showed the highest accuracy when stored in a nearest neighbor classifier and applied to the training set (Skalak 1994; Cameron-Jones 1995). We have previously shown that prototype sampling can effectively construct small, accurate nearest neighbor classifiers in some domains (Skalak 1994). The algorithms are different in only one respect. The Deliberate Misclassification algorithm deliberately modifies the target class labels of selected instances in the training set, while the Coarse Reclassification algorithm does not.

The name "Coarse Reclassification" is derived from the complementary classifier's salient characteristics: it classifies again the training data already classified by the base classifier, but does so with coarse hypotheses only. Breiman has shown that bootstrap aggregating (*bagging*) nearest neighbor classifiers in general will not be effective because they are *stable*: small perturbations in the training data will not change the hypothesis very much (Breiman 1994a). Therefore we have adopted a policy of what might be called *radical destabilization* of the nearest neighbor algorithm, enforced by choosing a very small number of prototypes. This artificial injection of instability yields a less stable classifier that nevertheless may be usefully included as a level-0 component in a stacked generalizer.

The Deliberate Misclassification algorithm is heuristic and is based on a simple intuition. If the goal is to create a complementary classifier that correctly classifies instances that the base classifier has misclassified, then the base classifier decision boundaries must be altered by the complementary classifier. This algorithm moves a decision boundary by changing the target classes of selected instances in a neighborhood of the boundary, as suggested by Figure 2.

Specifically, the Deliberate Misclassification algorithm modifies the class targets of a subset of the in-

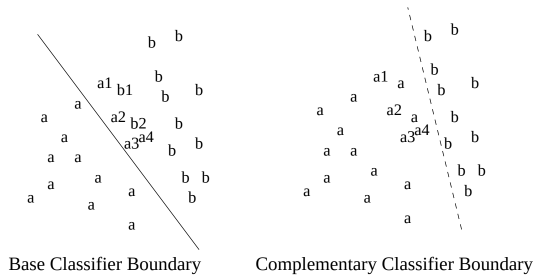

Figure 2: Shifting a decision boundary by Deliberate Misclassification. Instances to the left[right] of the boundary are predicted to be from class A [B]. Instances labeled a [b] are from class A [B]. The errors of the base classifier are labeled $a_i, i = 1, \ldots, 3$. Deliberate Misclassification changes the class targets of instances denoted $b_j, j = 1, 2$ from b to a in the training set for the complementary classifier.

stances that are in a neighborhood (of cardinality k for a user-specified parameter k) of the errors made by the base classifier along the most *asymmetric* inter-class boundaries. The degree of asymmetry of a boundary (A, B) between classes A and B may be characterized by the difference between the number of errors made by the base classifier in one direction (*e.g.,* instances of class A misclassified as members of class B) and the number of errors in the opposite direction (*e.g.,* B 's misclassified as A 's). More formally, if we let E be the confusion matrix for the base classifier on the training set, a maximally asymmetric inter-class boundary (A, B) is given by:

$$argmax_{(X,Y) \in \Pi \times \Pi}[E(X,Y) - E(Y,X)]$$

where Π is the set of class labels.

Instances near an asymmetric boundary are retargeted with the true class of the instances for which the base classifier made mistakes. This sleight-of-hand creates a more class-homogeneous neighborhood around the errors. The intuition is that since we employ instance-based classifiers that make predictions based on local class assignments, the base classifier errors are more likely to be correctly classified by the complementary classifier: now they live in a neighborhood of instances with the same class.

While not guaranteeing that only "helpful" instances will have their targets changed, altering class assignments near boundaries where the errors are asymmetric appears to be a useful heuristic (Skalak 1996). By concentrating on asymmetric boundaries, the algorithm raises the possibility that a complementary classifier may correctly classify a larger number of instances at the expense of misclassifying a smaller number, resulting in a net gain in accuracy. Nevertheless, there is unhedged risk that changing class labels will introduce more errors than it removes. Of course, the ultimate goal is not to create an accurate complementary classifier, but to create one that contributes to a more accurate stacked generalizer.

To describe the algorithms compactly, pseudocode is given in Figure 3. These algorithms are two points in a family of algorithms, indexed by k, the cardinality of the neighborhood of an error made by the base classifier in which targets may be changed. Coarse Reclassification corresponds to $k = 0$; the Deliberate Misclassification algorithm evaluated in this study uses $k = 1$. Variations on Deliberate Misclassification may make multiple calls to `modify-targets` to change the targets of instances along a user-specified number of asymmetric class boundaries.

In the call to `train-stacked-classifier` in Figure 3, leave-one-out cross validation was used to train the level-1 classifier. Each level-0 instance was withheld in turn from the training set for the level-0 component classifiers, the level-0 classifiers were trained, and a prediction was made for the class of the withheld instance by each of the component classifiers. Each vector of predictions was then assigned the class of the withheld instance. The collection of such vectors formed the training set for the level-1 classifier.

The nearest neighbor algorithm used by the level-0 classifiers defined the distance between two instances to be an unweighted sum of feature-by-feature differences on symbolic and numeric features. For symbolic features, if the attribute values were identical, their difference was 0; otherwise, their difference was 1. The distance between two instances on a real-valued attribute was the absolute value of the difference between scaled values. (Real values were linearly scaled and outlying values trimmed to fit within $[0.0, 1.0]$, where 0.50 represented the mean value for an attribute across all training instances, 0.0 represented values three or more standard deviations less than the mean, and 1.0 represented values three or more standard deviations greater than the mean.) The ID3 level-1 combining algorithm used the gain ratio metric to select branching features.

Our study is controlled to isolate the effect of changing the class labels. We show that this difference does not greatly affect the generalization accuracy of the resulting stacked generalizer, but that it does increase the diversity of the complementary classifier, and for some data sets changes the way that the accuracy of the base nearest neighbor classifier is boosted.

Key:
C_0 : `level-0 base classifier`
C_1 : `level-0 complementary classifier`
L_1 : `level-1 classifier`
T : `training set`
T' : `training set with (possibly) modified class labels`
$T_E \subset T$: `instances whose class labels will be changed`
k : `neighborhood cardinality`

procedure `boost-by-prototype-sampling` (C_0, T, L_1, k)
$T' \leftarrow T$
$C_0 \leftarrow$ `train` (C_0, T)
$T' \leftarrow$ `modify-targets` (C_0, T, k)
  ;;randomly sample prototypes
$C_1 \leftarrow$ `train-complementary` (C_1, T')
$S \leftarrow$ `create-stacked-classifier` (L_1, C_0, C_1)
$S \leftarrow$ `train-stacked-classifier` (S, T)

procedure `modify-targets` (C_0, T, k)
$T_E \leftarrow \phi$
$E(A, B) \leftarrow$ *heuristically identify training instances from one class (A) that* C_0 *misclassifies as members of one other class (B) (see text)*
`for each` $a \in E(A, B)$
  `do` $U \leftarrow$ `k-nearest-neighbors` (a, T, k)
    `for each` $b \in U$
      `do if` $class(b) = B$
        `then`
          $class(b) \leftarrow A$
          $T_E \leftarrow T_E \cup \{b\}$
$T' \leftarrow T_E \cup (T - T_E)$

Coarse Reclassification:
`boost-by-prototype-sampling` $(C_0, T, L_1, 0)$
Deliberate Misclassification:
`boost-by-prototype-sampling` $(C_0, T, L_1, 1)$

Figure 3: Pseudocode for the boost-by-prototype-sampling family indexed by k, the cardinality of a neighborhood in which class labels may be changed.

Experiments

Our experimental objective is to compare the base classifier with the stacked generalizer created by each of the Coarse Reclassification and Deliberate Misclassification algorithms and with the level-0 classifiers created by those algorithms. We ran these two algorithms on 13 data sets, selected mostly from the representative data sets for the U.C.I. Repository identified by Zheng (Zheng 1993). Some characteristics of the training data are given in Table 1. A one-tail, matched-pair

Table 1: Characteristics of the training data. Data sets in boldface have the four lowest values of the average number of training instances per class.

Data	*Training Instances*	*Classes*	*Ave. Instances/Class*
Breast Cancer Ljubljana	252	2	126.0
Breast Cancer Wisconsin	621	2	310.5
Heart Disease	270	2	135.0
Diabetes	684	2	342.0
Glass Recognition	189	6	**31.5**
Hepatitis	135	2	67.5
Iris Plants	135	3	45.0
LED-24 Digit	180	10	**18.0**
LED-7 Digit	450	10	45.0
Lymphography	126	4	**31.5**
Monks 2	387	2	193.5
Promoter	90	2	45.0
Soybean	270	19	**14.2**

t-test was used for significance testing of the stacked generalizer accuracy against the base classifier, using 10-fold cross validation with equal-sized partition segments. Class targets were changed along two inter-class boundaries by Deliberate Misclassification.

We explore a number of subsidiary questions.

How well do these algorithms boost? Table 2 shows that Deliberate Misclassification statistically significantly increases classification accuracy on seven data sets ($p \leq 0.10$) and Coarse Reclassification significantly increases classification accuracy on six data sets. (The algorithms boost nominal accuracy on the same nine data sets.) This result may be surprising in itself, given that the algorithms employ a single complementary classifier, the complementary classifier is restricted to a high-bias model class, and the algorithms rely on a small number of random samples to select prototypes for the complementary level-0 classifier.

On two data sets (Breast Cancer Ljubljana and Monks 2), the base nearest neighbor displays lower accuracy than achieved by always predicting the most frequent class, although on these two data sets the complementary classifier also usually failed to beat this default.

When and why do these algorithms fail to work? Both algorithms construct stacked generalizers that are statistically significantly worse than the base classifier on four data sets: Glass Recognition, LED-24 Digit, Lymphography and Soybean. These negative results may be judged against Wolpert's observation that no generalizing scheme will always improve accuracy (Wolpert 1992, p.242). Referring to Table 1, we see that these four data sets have the lowest average number of training instances per class. One possible explanation is poverty of data: where the number of instances per class is low, there are not sufficiently many data available to characterize a class, and random sampling of prototypes will not work well to produce components that may combined successfully.

Does Deliberate Misclassification's modification of selected training labels affect generalization accuracy? Deliberate Misclassification and Coarse Reclassification show statistically comparable accuracy on nine data sets. At the $p \leq 0.05$ significance level, Deliberate Misclassification is better than Coarse Reclassification on two data sets (Heart Disease and Soybean), but Coarse Reclassification is better on two (Breast Cancer Ljubljana and Diabetes).

Is there a difference in how the two algorithms boost accuracy? Restricting our attention to the nine data sets on which the Coarse Reclassification algorithm significantly boosted accuracy ($p \leq 0.10$) or failed to boost accuracy ($p \geq 0.90$), we observe from Table 2 that its success or failure was determined solely by whether the algorithm created a complementary level-0 classifier that was more accurate or less accurate than the base classifier. (This observation holds for nominal accuracy on 12 of 13 data sets — Promoter is the exception.)

On the other hand, Deliberate Misclassification demonstrates that it is possible to boost the accuracy of a base classifier by combining it with a *less* accurate classifier. Table 2 shows that for two data sets (Breast Cancer Ljubljana and Monks 2) this algorithm creates a complementary classifier that is less accurate than the base classifier, but still creates a statistically significantly better stacked generalizer ($p \leq 0.10$). (If we do not restrict ourselves to this level of significance, we see that in the four of nine data sets on which Deliberate Misclassification boosted accuracy, it did so by creating a less accurate component.) Whereas the success or failure of Coarse Reclassification is determined entirely by the relative accuracy of the level-0 component it creates, Deliberate Misclassification demonstrates the capacity to boost a classifier by pairing it with one that is less accurate.

Are the level-0 classifiers created by Deliberate Misclassification more diverse than those created by Coarse

Table 2: Average percent correct on test data for the level-0 complementary classifiers and stacked generalizers created by the Coarse Reclassification (CR) and Deliberate Misclassification (DM) algorithms. Significance levels (p-values) are given in parentheses for the stacked generalizers with respect to the base classifier.

Data	*Base-NN*	*Level-0-CR*	*Level-0-DM*	*Stacked-CR*	*Stacked-DM*
Breast Cancer Ljubljana	67.5	72.8	66.8	75.4 (0.006)	72.1 (0.084)
Breast Cancer Wisconsin	96.2	96.8	97.1	97.8 (0.016)	97.4 (0.076)
Heart Disease	77.0	78.0	80.7	78.0 (0.371)	81.3 (0.038)
Diabetes	68.7	73.7	70.7	73.7 (0.005)	71.8 (0.055)
Glass Recognition	73.3	51.9	52.4	71.9 (0.783)	70.5 (0.870)
Hepatitis	61.3	69.3	56.0	67.3 (0.091)	62.7 (0.254)
Iris Plants	91.3	94.7	96.0	93.3 (0.140)	94.7 (0.026)
LED-24 Digit	36.5	35.0	26.5	28.0 (0.995)	30.5 (0.937)
LED-7 Digit	71.2	72.6	72.8	73.2 (0.021)	73.2 (0.048)
Lymphography	80.7	77.9	72.1	75.0 (0.948)	76.4 (0.903)
Monks 2	48.8	50.9	42.1	66.7 (0.000)	65.6 (0.000)
Promoter	79.0	61.0	70.0	80.0 (0.379)	82.0 (0.197)
Soybean	91.0	63.7	67.7	86.0 (0.999)	89.3 (0.931)

Reclassification? We define diversity as the percent of test instances for which the base and complementary classifiers make different predictions but for which one of them is correct (*cf.* (Opitz and Shavlik 1995; Ali and Pazzani 1995; Krogh and Vedelsby 1995)). From Table 3 it can be seen that Deliberate Misclassification creates a level-0 complementary classifier that is at least as diverse as that created by Coarse Reclassification for 11 of the 13 data sets.

Further, the three data sets for which Deliberate Misclassification shows the greatest increase in level-0 diversity (Breast Cancer Ljubljana, Hepatitis and Monks 2) are those for which the Deliberate Misclassification algorithm boosts base-classifier accuracy by creating a less accurate complementary classifier (although significance for Hepatitis is only at the 0.25 level.) In effect, to accomplish boosting on those three data sets, Deliberate Misclassification increased component diversity to overcome a deficit in accuracy.

Table 3: Diversity (%) of the complementary level-0 classifier created by the two algorithms, and the percent increase in diversity of Deliberate Misclassification (DM) over Coarse Reclassification (CR).

Data	*CR*	*DM*	*Inc.*
Breast Cancer Ljubljana	23.9	32.1	**34.3**
Breast Cancer Wisconsin	3.2	3.2	0.0
Heart Disease	18.7	19.7	5.3
Diabetes	28.6	30.9	8.3
Glass Recognition	35.7	37.1	4.0
Hepatitis	32.0	38.7	**20.8**
Iris Plants	6.7	6.7	0.0
LED-24 Digit	39.5	40.0	1.3
LED-7 Digit	13.4	14.4	7.5
Lymphography	28.6	30.0	5.0
Monks 2	44.9	54.2	**20.7**
Promoter	42.0	31.0	-26.0
Soybean	29.3	29.0	-1.1

Summary

We have introduced two boosting algorithms that effectively boost accuracy by creating only one additional efficient, high-bias nearest neighbor classifier. Despite the fact that both algorithms construct classifiers from the same model class using the same stochastic techniques, they apparently owe their success on some data sets to having exploited different combinations of accuracy and diversity. The Coarse Reclassification algorithm derived its complementary power from constructing a more accurate level-0 classifier than the base. Deliberate Misclassification generally created a more diverse level-0 classifier, and demonstrated the capacity to create a less accurate — but more diverse — complementary classifier. These algorithms are two representatives of an indexed family of algorithms that demonstrate varying degrees of diversity and accuracy. Our analysis may encourage the development of families of boosting algorithms that can tailor the combination of those two factors according to the domain and task at hand.

Acknowledgments

This work was supported by a University of Massachusetts Graduate Fellowship and computing facilities provided by the Cornell University Department of Computer Science.

References

Ali, K.M. and Pazzani, M.J. 1995. Error reduction through learning multiple descriptions. *Machine Learning* (To appear). Also published as Dept. of Information and Computer Science Technical Report 95-39, University of California, Irvine, CA.

Battiti, R. and Colla, A.M. 1994. Democracy in Neural Nets: Voting Schemes for Classification. *Neural Networks* 7:691–707.

Breiman, L. 1992. Stacked Regressions. Technical Report 367, Department of Statistics, University of California, Berkeley, CA.

Breiman, L. 1994a. Bagging predictors. Technical Report 421, Department of Statistics, University of California, Berkeley, CA.

Breiman, L. 1994b. NIPS*94 Tutorial, Statistics and Nets: Understanding Nonlinear Models from Their Linear Relatives. Neural Information Processing Systems, 1994, Denver, CO.

Cameron-Jones, M. 1995. Instance Selection by Encoding Length Heuristic with Random Mutation Hill Climbing. In *Proceedings of the Eighth Australian Joint Conference on Artificial Intelligence.* World Scientific. 99–106.

Chan, P.K. and Stolfo, S.J. 1993a. Experiments on multistrategy learning by meta-learning. In *Proceedings of the Second International Conference on Information and Knowledge Management.* 314–323.

Chan, P.K. and Stolfo, S.J. 1993b. Toward Parallel and Distributed Learning by Meta-Learning. In *Working Notes, AAAI Workshop on Knowledge Discovery in Databases*, San Mateo, CA. AAAI Press/MIT Press. 227–240.

Chan, P.K. and Stolfo, S.J. 1995. A Comparative Evaluation of Voting and Meta-Learning on Partitioned Data. In *Proceedings of the Twelfth International Conference on Machine Learning*, San Mateo, CA. Morgan Kaufmann. 90–98.

Drucker, H. and Cortes, C. 1996. Boosting Decision Trees. In *Advances in Neural Information Processing Systems, 8.* MIT Press, Las Vegas, NV.

Drucker, H.; Cortes, C.; Jackel, L.D.; LeCun, Y.; and Vapnik, V. 1994. Boosting and other ensemble methods. *Neural Computation* 6(6):1289–1301.

Freund, Y. and Schapire, R.E. 1995. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. In *Proceedings of the Second European Conference on Computational Learning Theory.* Barcelona, Spain.

Hansen, L.K. and Salamon, P. 1990. Neural Network Ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12:993–1001.

Jacobs, R. A.; Jordan, M. I.; and Barto, A. G. 1991. Task Decomposition through Competition in a Modular Connectionist Architecture: The What and Where Vision Tasks. *Cognitive Science* 15:219–250.

Kong, E.B. and Dietterich, T.G. 1995. Error-correcting output coding corrects bias and variance. In *Proceedings of the Twelfth International Conference on Machine Learning.* Morgan Kaufmann, San Francisco, CA. 313–321.

Krogh, A. and Vedelsby, J. 1995. Neural Network Ensembles, Cross Validation and Active Learning. In Tesauro, G.; Touretzky, D.; and Leen, T., editors, *Advances in Neural Information Processing Systems, 7.* MIT Press.

Maclin, R. and Shavlik, J.W. 1995. Combining the predictions of multiple classifiers: Using competitive learning to initialize neural networks. In Mellish, C.S., editor, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence.* Morgan Kaufmann, San Mateo, CA. 524–530.

Opitz, D.W. and Shavlik, J.W. 1995. Generating Accurate and Diverse Members of a Neural-Network Ensemble. In Mozer, M.C. and Hasselmo, M.E., editors, *Advances in Neural Information Processing Systems, 8.* MIT Press.

Schapire, R.E. 1990. The Strength of Weak Learnability. *Machine Learning* 5:197–227.

Selfridge, O. G. 1959. Pandemonium: A Paradigm for Learning. In *Proceedings of the Symposium on the Mechanization of Thought Processes*, Teddington, England. National Physical Laboratory, H.M. Stationery Office, London. 511–529.

Skalak, D. B. 1994. Prototype and Feature Selection by Sampling and Random Mutation Hill Climbing Algorithms. In *Proceedings of the Eleventh International Conference on Machine Learning*, New Brunswick, NJ. Morgan Kaufmann. 293–301.

Skalak, D.B. 1995. Prototype selection for composite nearest neighbor classifiers. Technical Report 95-74, Department of Computer Science, University of Massachusetts, Amherst, MA.

Skalak, D.B. 1996. *Prototype Selection for Composite Nearest Neighbor Classifiers.* Ph.D. Dissertation, Dept. of Computer Science, University of Massachusetts, Amherst, MA (forthcoming).

Wolpert, D. 1992. Stacked Generalization. *Neural Networks* 5:241–259.

Zhang, X.; Mersirov, J.P.; and D.L.Waltz, 1992. A Hybrid System for Protein Secondary Structure Prediction. *Journal of Molecular Biology* 225:1049–1063.

Zheng, Z. 1993. A Benchmark for Classifier Learning. In *Proceedings of the Sixth Australian Joint Conference on Artificial Intelligence.* World Scientific Publisher (An extended version is available as Technical Report TR474, Basser Department of Computer Science, The University of Sydney, available by anonymous ftp to ftp.cs.su.oz.au in the /pub/tr directory). 281–286.