Ensemble Learning

Martin Sewell
Department of Computer Science
University College London

April 2007 (revised August 2008)

1 Introduction

The idea of *ensemble learning* is to employ multiple learners and combine their predictions. There is no definitive taxonomy. Jain, Duin and Mao (2000) list eighteen classifier combination schemes; Witten and Frank (2000) detail four methods of combining multiple models: bagging, boosting, stacking and error-correcting output codes whilst Alpaydin (2004) covers seven methods of combining multiple learners: voting, error-correcting output codes, bagging, boosting, mixtures of experts, stacked generalization and cascading. Here, the literature in general is reviewed, with, where possible, an emphasis on both theory and practical advice, then the taxonomy from Jain, Duin and Mao (2000) is provided, and finally four ensemble methods are focussed on: bagging, boosting (including AdaBoost), stacked generalization and the random subspace method.

2 Literature Review

Wittner and Denker (1988) discussed strategies for teaching layered neural networks classification tasks.

Schapire (1990) introduced *boosting* (see Section 5 (page 9)). A theoretical paper by Kleinberg (1990) introduced a general method for separating points in multidimensional spaces through the use of stochastic processes called *stochastic discrimination* (SD). The method basically takes poor solutions as an input and creates good solutions. Stochastic discrimination looks promising, and later led to the random subspace method (Ho 1998). Hansen and Salamon (1990) showed the benefits of invoking ensembles of similar neural networks.

Wolpert (1992) introduced *stacked generalization*, a scheme for minimizing the generalization error rate of one or more generalizers (see Section 6 (page 10)).Xu, Krzyżak and Suen (1992) considered methods of combining multiple classifiers and their applications to handwriting recognition. They claim that according to the levels of output information by various classifiers, the problems of combining multiclassifiers can be divided into three types. They go

on to compare three approaches from the first type: voting, Bayesian formalism and Dempster-Shafer formalism. They found that the performance of individual classifiers could be improved significantly and, if forced to pick one, they'd recommend the Dempster-Shafer formalism since it can obtain high recognition and reliability rates simultaneously and robustly.

Perrone and Cooper (1993) presented a general theoretical framework for ensemble methods of constructing significantly improved regression estimates. Jordan and Jacobs (1993) presented a hierarchical mixtures of experts model.

Ho, Hull and Srihari (1994) suggest a multiple classifier system based on rankings. In the field of handwritten digit recognition, Battiti and Colla (1994) found that the use of a small number of neural nets (two to three) with a sufficiently small correlation in their mistakes reaches a combined performance that is significantly higher than the best obtainable from the individual nets.

Cho and Kim (1995) combined the results from multiple neural networks using fuzzy logic which resulted in more accurate classification. Bishop (1995) covers the theoretical aspects of committees of neural networks. If L networks produce errors which have zero mean and are uncorrelated, then the sum-of-squares error can be reduced by a factor of L simply by averaging the predictions of the L networks. Although in practice the errors are likely to be highly correlated. However, by making use of Cauchy's inequality, he also shows that the committee averaging process cannot produce an increase in the expected error. It is important to note that because the reduction in error can be viewed as arising from reduced variance due to the averaging over many solutions, when individual members are selected, the optimal trade-off between bias and variance should have relatively smaller bias, since the extra variance can be removed by averaging. If greater weight is given to the committee members that make the better predictions, the error can be reduced further. The benefits of committee averaging are not limited to sum-of-squares error, but apply to any error function which is convex. Bishop also shows that the concept of a committee arises naturally in a Bayesian framework. Krogh and Vedelsby (1995) showed that there is a lot to be gained from using unlabeled data when training ensembles. Lam and Suen (1995) studied the performance of four combination methods: the majority vote, two Bayesian formulations and a weighted majority vote (with weights obtained through a genetic algorithm). They conclude: 'in the absence of a truly representative training set, simple majority vote remains the easiest and most reliable solution among the ones studied here.'

Tumer and Ghosh (1996) showed that combining neural networks linearly in output space reduces the variance of the actual decision region boundaries around the optimum boundary. Of great practical importance, Sollich and Krogh (1996) found that in large ensembles, it is advantageous to use under-regularized students, which actually over-fit the training data. This allows one to maximize the benefits of the variance-reducing effects of ensemble learning. Freund and Schapire (1996) introduced AdaBoost (see Section 5.1 (page 9)). Breiman (1996) introduced bagging (see Section 4 (page 9)).

Raftery, Madigan and Hoeting (1997) consider the problem of accounting for model uncertainty in linear regression models and offer two extensions to

Bayesian model averaging: Occam's window and Markov chain Monte Carlo. Woods, Kegelmeyer and Bowyer (1997) presented a method for combining classifiers that uses estimates of each individual classifier's local accuracy in small regions of feature space surrounding an unknown test sample. An empirical evaluation showed that their local accuracy approach was more effective than the classifier rank algorithm, the modified classifier rank algorithm and the behaviour-knowledge space (BKS) algorithm (which performed worst). In fact, on average, the classifier rank algorithm and the BKS algorithm both failed to outperform the single best classifier. The authors of the article believe that the combining of classifiers works best with large data sets with data distributions that are too complex for most individual classifiers. Larkey and Croft (1997) found that combining classifiers in text categorization improved performance. Lam and Suen (1997) analysed the application of majority voting to pattern recognition.

Kittler (1998) developed a theoretical framework for combining classifiers in the two main fusion scenarios: fusion of opinions based on identical and on distinct representations. For the shared representation they showed that here the aim of fusion was to obtain a better estimation of the appropriate a posteriori class probabilities. For distinct representations they pointed out that the techniques based on the benevolent sum-rule fusion are more resilient to errors than those derived from the severe product rule. In both cases (distinct and shared representations), the expert fusion involves the computation of a linear or non-linear function of the a posteriori class probabilities estimated by the individual experts. Kittler, *et al.* (1998) developed a common theoretical framework for classifier combination. An experimental comparison between the product rule, sum rule, min rule, max rule, median rule and majority voting found that the sum rule outperformed other classifier combinations schemes, and sensitivity analysis showed that the sum rule is most resilient to estimation errors (which may be a plausible explanation for its superior performance). Ho (1998) introduced the random subspace method for constructing decision forests (see Section 7 (page 10)). The method worked well in practice and was shown to perform best when the dataset has a large number of features and not too few samples. Schapire, *et al.* (1998) offer an explanation for the effectiveness of voting methods. They show that this phenomenon is related to the distribution of margins of the training examples with respect to the generated voting classification rule, where the margin of an example is simply the difference between the number of correct votes and the maximum number of votes received by any incorrect label.

Schapire (1999) introduces the boosting algorithm AdaBoost, and explains the underlying theory of boosting. Opitz and Maclin (1999) compared bagging and two boosting methods: AdaBoost and arching. They found that in a low noise regime, boosting outperforms bagging, which outperforms a single classifier, whilst as a general technique bagging is the most appropriate. Opitz (1999) consider feature selection for ensembles. Miller and Yan (1999) developed a critic-driven ensemble for classification. Hoeting, *et al.* (1999) provide a tutorial on Bayesian model averaging (BMA). Liu and Yao (1999) presented

negative correlation learning for neural network ensembles.

Jain, Duin and Mao (2000) include a section on classifier combination. They list reasons for combining multiple classifiers: one may have different feature sets, different training sets, different classification methods or different training sessions, all resulting in a set of classifiers whose results may be combined with the hope of improved overall classification accuracy. They provide a taxonomy, see Table 1 (page 7). In terms of experimental work, they train twelve classifiers on six feature sets from a digit dataset and use four methods of classifier combination—median, product, nearest mean and 1-NN—across both the different feature sets and the different classifiers. Measuring performance against the best single result, my own conclusions from their results are that 1) there is no benefit in just combining different classifiers across the same feature set and 2) there is substantial benefit in combining the results of one classifier across different feature sets (1-NN worked best, but voting failed). However, when the classifiers are first combined on one feature set at a time, and then these results are combined, then using the nearest mean method for both stages of model combination gave the best overall result. This was also the best result of the entire experiment. Kleinberg (2000) bridged the gap between the theoretical promise shown by stochastic discrimination and a practical solution by providing the algorithmic implementation. He also showed that stochastic discrimination outperformed both boosting and bagging in the majority of benchmark problems that it was tested on. Kuncheva, *et al.* (2000) consider whether independence is good for combining classifiers. Their results support the intuition that negatively related classifiers are better than independent classifiers, and they also show that this relationship is ambivalent. Dietterich (2000) compared the effectiveness of randomization, bagging and boosting for improving the performance of the decision-tree algorithm C4.5. Their experiments showed that in situations with little or no classification noise, randomization is competitive with (and perhaps slightly superior to) bagging but not as accurate as boosting. In situations with substantial classification noise, bagging is much better than boosting, and sometimes better than randomization. Kuncheva and Jain (2000) designed two classifier fusion systems using genetic algorithms and found that selection of classifiers and (possibly overlapping) feature subsets worked well, but selection of disjoint feature subsets did not. Tax, *et al.* (2000) sought to answer the question of whether to combine multiple classifiers by averaging or multiplying. They concluded that averaging-estimated posterior probabilities is to be preferred in the case when posterior probabilities are not well estimated. Only in the case of problems involving multiple classes with good estimates of posterior class probabilities did the product combination rule outperform the mean combination rule. Liu, Yao and Higuchi (2000) presented evolutionary ensembles with negative correlation learning (EENCL). Allwein, Schapire and Singer (2000) proved a general empirical multiclass loss bound given the empirical loss of the individual binary learning algorithms.

In a PhD thesis, Skurichina (2001) tackled the problem of stabilizing weak classifiers and compares bagging, boosting and the random subspace method. Bagging is useful for weak and unstable classifiers with a non-decreasing learning

curve and critical training sample sizes. Boosting is beneficial only for weak, simple classifiers, with a non-decreasing learning curve, constructed on large training sample sizes. The random subspace method is advantageous for weak and unstable classifiers that have a decreasing learning curve and are constructed on small and critical training sample sizes.

Kuncheva (2002a) give formulas for the classification error for the following fusion methods: average, minimum, maximum, median, majority vote and oracle. For a uniformly distributed posterior probability, the minimum/maximum method performed the best; whilst for normally distributed errors, the fusion methods all gave a very similar performance. Kuncheva (2002b) presented a combination of classifier selection and fusion by using statistical inference to switch between the two. In their experiments, there was no clear preference of one combination approach over the rest, the only consistent pattern being that the improvement over the best individual classifier was negligible. Shipp and Kuncheva (2002) studied the relationships between different methods of classifier combination and measures of diversity in combining classifiers. The only positive correlation was that the 'double-fault measure' of diversity and the measure of difficulty both showed reasonable correlation with majority vote and naive Bayes combinations (a not unexpected result). The ambiguous relationship between diversity and accuracy discourages optimising the diversity. Skurichina and Duin (2002) applied and compared bagging, boosting and the random subspace method to linear discriminant analysis. They discovered that boosting is useful for large training sample sizes, whilst bagging and the random subspace method are useful for critical training sample sizes. In a very good paper, Valentini and Masulli (2002) present an overview of ensemble methods. Dietterich (2002) published a review of ensemble learning.

Kittler and Alkoot (2003) investigated the 'sum' versus 'majority vote' in multiple classifier systems. They showed that for Gaussian estimation error distributions, sum always outperforms vote; whilst for heavy tail distributions, vote may outperform sum. This is of especial interest to the financial domain with the presence of leptokurtosis in market returns. Kuncheva, *et al.* (2003) derived upper and lower limits on the majority vote accuracy for individual classifiers. They deduce that negative pairwise dependence between classifiers is best, and ideally all pairs of classifiers in the pool should have the same negative dependence. They also deduce that diversity is not always beneficial. Kuncheva and Whitaker (2003) considered measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Their results raise some doubts about the usefulness of diversity measures in building classifier ensembles in real-life pattern recognition problems. Topchy, Jain and Punch (2003) investigate clustering ensembles.

Džeroski and Ženko (2004) considered the construction of ensembles of heterogeneous classifiers using stacking and showed that they perform (at best) comparably to selecting the best classifier from the ensemble by cross-validation. They also proposed two new methods for stacking by extending the method with probability distributions and multiresponse linear regression. They showed that the latter extension performs better than existing stacking approaches and bet-

ter than selecting the best classifier by cross-validation. Chawla, *et al.* (2004) proposed a framework for building hundreds or thousands of classifiers on small subsets of data in a distributed environment. Their experiments showed that their approach is fast, accurate and scalable. In a relevant and interesting paper, Evgeniou, Pontil and Elisseeff (2004) studied the leave-one-out and generalization errors of ensembles of kernel machines such as SVMs. They found that the best SVM and the best ensembles had about the same test performance: 'with appropriate tuning of the parameters of the machines, combining SVMs does not lead to performance improvement compared to a single SVM.' However, ensembles of kernel machines are more stable learning algorithms than the equivalent single kernel machine, i.e. bagging increases the stability of unstable learning machines. Topchy, Jain and Punch (2004) proposed a solution to the problem of clustering combination by offering a probabilistic model of consensus using a finite mixture of multinomial distributions in a space of clusterings. A combined partition is found as a solution to the corresponding maximum likelihood problem using the expectation-maximization (EM) algorithm. In a directly relevant paper, Valentini and Dietterich (2004) analysed bias-variance in SVMs for the development of SVM-based ensemble methods. They suggest two promising approaches for designing ensembles of SVMs. One approach is to employ low-bias SVMs as base learners in a bagged ensemble, whilst the other approach is to apply bias-variance analysis to construct a heterogeneous, diverse set of accurate and low-bias classifiers.

In March 2005 the journal *Information Fusion* ran a special issue on 'Diversity in multiple classifier systems'; Ludmila I. Kuncheva gave a guest editorial (Kuncheva 2005). Melville and Mooney (2005) presented a new method for generating ensembles, DECORATE (Diverse Ensemble Creation by Oppositional Relabeling of Artificial Training Examples), that directly constructs diverse hypotheses using additional artificially-constructed training examples. Their approach consistently outperformed the base classifier, bagging and random forests; and outperformed AdaBoost on small training sets and achieved comparable performance on larger training sets. Ruta and Gabrys (2005) provide a revision of the classifier selection methodology and evaluate the practical applicability of diversity measures in the context of combining classifiers by majority voting. Fumera and Roli (2005) presented a theoretical and experimental analysis of linear combiners for classifier fusion. Their theoretical analysis shows how the performance of linear combiners depends on the performance of individual classifiers, and on the correlation between their outputs. In particular, they considered the improvements gained from using a weighted average over the simple average combining rule. García-Pedrajas, Hervás-Martnez and Ortiz-Boyer (2005) present a cooperative coevolutive approach for designing neural network ensembles.

Chandra and Yao (2006) used an evolutionary framework to evolve hybrid ensembles. The framework treats diversity and accuracy as evolutionary pressures which are exerted at multiple levels of abstraction. Their method was shown to be effective. Reyzin and Schapire (2006) show that boosting the margin can also boost classifier complexity. They conclude that maximizing the

margins is desirable, but not necessarily at the expense of other factors, especially base-classifier complexity. Hadjitodorov, Kuncheva and Todorova (2006) found that ensembles which exhibited a moderate level of diversity produced better cluster ensembles. Kuncheva and Vetrov (2006) evaluated the stability of k-means cluster ensembles with respect to random initialization. They found that ensembles are generally more stable, and that the relationship between stability and accuracy with respect to the number of clusters strongly depends on the data set. They also created a new combined stability index, the sum of the pairwise individual and ensemble stabilities, which was effective.

Canuto, *et al.* (2007) investigated how the choice of component classifiers can affect the performance of several combination methods (selection-based and fusion-based methods). One key result was that the highest accuracies were almost always reached by using hybrid structures. Kuncheva and Rodríguez (2007) proposed a combined fusion-selection approach to classifier ensemble design, which they called the 'random linear oracle'. Each classifier in the ensemble is replaced by a miniensemble of a pair of subclassifiers with a random linear oracle (in the form of a hyperplane) to choose between the two. Experiments showed that all ensemble methods benefited from their approach. Hansen (2007) considers the problem of selection of weights for averaging across least squares estimates obtained from a set of models and proposes selecting the weights by minimizing a Mallows criterion. Bühlmann and Hothorn (2007) present a statistical perspective on boosting. They give an overview on theoretical concepts of boosting as an algorithm for fitting statistical models, and also look at the methodology from a practical point of view.

Zhang and Zhang (2008) propose a local boosting algorithm, based on the boosting-by-resampling version of AdaBoost, for dealing with classification. Their experimental results found the algorithm to be more accurate and robust than AdaBoost. Claeskens and Hjort (2008) publish *Model Selection and Model Averaging*. The book explains, discusses and compares model choice criteria, including the AIC, BIC, DIC and FIC. Leap, *et al.* (2008) investigated the effects of correlation and autocorrelation on classifier fusion and optimal classifier ensembles. Results included the finding that fusion methods employing neural networks outperformed those methods that fuse based on Boolean rules.

3 Taxonomy

Table 1 provides a taxonomy of ensemble methods which was taken from Jain, Duin and Mao (2000).

Table 1: Ensemble methods

Scheme	Architecture	Trainable	Adaptive	Info-level	Comments
Voting	Parallel	No	No	Abstract	Assumes independent classifiers

Continued on next page

Scheme	Architecture	Trainable	Adaptive	Info-level	Comments
Sum, mean, median	Parallel	No	No	Confidence	Robust; assumes independent confidence estimators
Product, min, max	Parallel	No	No	Confidence	Assumes independent features
Generalized ensemble	Parallel	Yes	No	Confidence	Considers error correlation
Adaptive weighting	Parallel	Yes	Yes	Confidence	Explores local expertise
Stacking	Parallel	Yes	No	Confidence	Good utilization of training data
Borda count	Parallel	Yes	No	Rank	Converts ranks into confidences
Logistic regression	Parallel	Yes	No	Rank confidence	Converts ranks into confidences
Class set reduction	Parallel cascading	Yes/No	No	Rank confidence	Efficient
Dempster-Shafer	Parallel	Yes	No	Rank confidence	Fuses non-probabilistic confidences
Fuzzy integrals	Parallel	Yes	No	Confidence	Fuses non-probabilistic confidences
Mixture of local experts (MLE)	Gated parallel	Yes	Yes	Confidence	Explores local expertise; joint optimization
Hierarchical MLE	Gated parallel hierarchical	Yes	Yes	Confidence	Same as MLE; hierarchical
Associative switch	Parallel	Yes	Yes	Abstract	Same as MLE, but no joint optimization
Bagging	Parallel	Yes	No	Confidence	Needs many comparable classifiers
Boosting	Parallel hierarchical	Yes	No	Abstract	Improves margins; unlikely to overtrain, sensitive to mislabels; needs many comparable classifiers

Continued on next page

Scheme	Architecture	Trainable	Adaptive	Info-level	Comments
Random subspace	Parallel	Yes	No	Confidence	Needs many comparable classifiers
Neural trees	Hierarchical	Yes	No	Confidence	Handles large numbers of classes

4 Bagging

Bagging (Breiman 1996), a name derived from *bootstrap aggregation*, was the first effective method of ensemble learning and is one of the simplest methods of arching[1]. The meta-algorithm, which is a special case of model averaging, was originally designed for classification and is usually applied to decision tree models, but it can be used with any type of model for classification or regression. The method uses multiple versions of a training set by using the bootstrap, i.e. sampling with replacement. Each of these data sets is used to train a different model. The outputs of the models are combined by averaging (in the case of regression) or voting (in the case of classification) to create a single output. Bagging is only effective when using unstable (i.e. a small change in the training set can cause a significant change in the model) non-linear models.

5 Boosting (Including AdaBoost)

Boosting (Schapire 1990) is a meta-algorithm which can be viewed as a model averaging method. It is the most widely used ensemble method and one of the most powerful learning ideas introduced in the last twenty years. Originally designed for classification, it can also be profitably extended to regression. One first creates a 'weak' classifier, that is, it suffices that its accuracy on the training set is only slightly better than random guessing. A succession of models are built iteratively, each one being trained on a data set in which points misclassified (or, with regression, those poorly predicted) by the previous model are given more weight. Finally, all of the successive models are weighted according to their success and then the outputs are combined using voting (for classification) or averaging (for regression), thus creating a final model. The original boosting algorithm combined three weak learners to generate a strong learner.

5.1 AdaBoost

AdaBoost (Freund and Schapire 1996), short for 'adaptive boosting', is the most popular boosting algorithm. It uses the same training set over and over again (thus it need not be large) and can also combine an arbitrary number of base-learners.

[1] *Arching* (adaptive reweighting and combining) is a generic term that refers to reusing or selecting data in order to improve classification.

6 Stacked Generalization

Stacked generalization (or *stacking*) (Wolpert 1992) is a different way of combining multiple models, that introduces the concept of a meta learner. Although an attractive idea, it is less widely used than bagging and boosting. Unlike bagging and boosting, stacking may be (and normally is) used to combine models of different types. The procedure is as follows:

1. Split the training set into two disjoint sets.
2. Train several base learners on the first part.
3. Test the base learners on the second part.
4. Using the predictions from 3) as the inputs, and the correct responses as the outputs, train a higher level learner.

Note that steps 1) to 3) are the same as cross-validation, but instead of using a winner-takes-all approach, the base learners are combined, possibly non-linearly.

7 Random Subspace Method

The *random subspace method* (RSM) (Ho 1998) is a relatively recent method of combining models. Learning machines are trained on randomly chosen subspaces of the original input space (i.e. the training set is sampled in the feature space). The outputs of the models are then combined, usually by a simple majority vote.

References

ALLWEIN, Erin L., Robert E. SCHAPIRE, and Yoram SINGER, 2000. Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers. *Journal of Machine Learning Research*, **1**, 113–141.

BATTITI, Roberto, and Anna Maria COLLA, 1994. Democracy in Neural Nets: Voting Schemes for Classification. *Neural Networks*, **7**(4), 691–707.

BISHOP, Christopher M., 1995. *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

BREIMAN, Leo, 1996. Bagging Predictors. *Machine Learning*, **24**(2), 123–140.

BÜHLMANN, Peter, and Torsten HOTHORN, 2007. Boosting Algorithms: Regularization, Prediction and Model Fitting. *Statistical Science*, **22**(4), 477–505.

CANUTO, Anne M. P., *et al.*, 2007. Investigating the Influence of the Choice of the Ensemble Members in Accuracy and Diversity of Selection-Based and Fusion-Based Methods for Ensembles. *Pattern Recognition Letters*, **28**(4), 472–486.

CHANDRA, Arjun, and Xin YAO, 2006. Evolving Hybrid Ensembles of Learning Machines for Better Generalisation. *Neurocomputing*, **69**(7–9), 686–700.

CHAWLA, Nitesh V., *et al.*, 2004. Learning Ensembles from Bites: A Scalable and Accurate Approach. *Journal of Machine Learning Research*, **5**, 421–451.

CHO, Sung-Bae, and Jin H. KIM, 1995. Multiple Network Fusion Using Fuzzy Logic. *IEEE Transactions on Neural Networks*, **6**(2), 497–501.

CLAESKENS, Gerda, and Nils Lid HJORT, 2008. *Model Selection and Model Averaging*. Volume 27 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge: Cambridge University Press.

DIETTERICH, Thomas G., 2000. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning*, **40**(2), 139–157.

DIETTERICH, Thomas G., 2002. Ensemble learning. *In*: Michael A. ARBIB, ed. *The Handbook of Brain Theory and Neural Networks*. Second ed., Bradford Books. Cambridge, MA: The MIT Press, pp. 405–408.

DŽEROSKI, Saso, and Bernard ŽENKO, 2004. Is Combining Classifiers with Stacking Better than Selecting the Best One? *Machine Learning*, **54**(3), 255–273.

EVGENIOU, Theodoros, Massimiliano PONTIL, and André ELISSEEFF, 2004. Leave One Out Error, Stability, and Generalization of Voting Combinations of Classifiers. *Machine Learning*, **55**(1), 71–97.

FREUND, Yoav, and Robert E. SCHAPIRE, 1996. Experiments with a New Boosting Algorithm. *In*: Lorenza SAITTA, ed. *Machine Learning: Proceedings of the Thirteenth International Conference (ICML '96)*. San Francisco, CA: Morgan Kaufmann, pp. 148–156.

FUMERA, Giorgio, and Fabio ROLI, 2005. A Theoretical and Experimental Analysis of Linear Combiners for Multiple Classifier Systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**(6), 942–956.

GARCÍA-PEDRAJAS, Nicolás, César HERVÁS-MARTNEZ, and Domingo ORTIZ-BOYER, 2005. Cooperative Coevolution of Artificial Neural Network Ensembles for Pattern Classification. *IEEE Transactions on Evolutionary Computation*, **9**(3), 271–302.

HADJITODOROV, Stefan T., Ludmila I. KUNCHEVA, and Ludmila P. TODOROVA, 2006. Moderate Diversity for Better Cluster Ensembles. *Information Fusion*, **7**(3), 264–275.

HANSEN, Bruce E., 2007. Least Squares Model Averaging. *Econometrica*, **75**(4), 1175–1189.

Hansen, Lars Kai, and Peter SALAMON, 1990. Neural Network Ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **12**(10), 993–1001.

Ho, Tin Kam, 1998. The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**(8), 832–844.

Ho, Tin Kam, Jonathan J. HULL, and Sargur N. SRIHARI, 1994. Decision Combination in Multiple Classifier Systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **16**(1), 66–75.

HOETING, Jennifer A., *et al.*, 1999. Bayesian Model Averaging: A Tutorial. *Statistical Science*, **14**(4), 382–401.

JAIN, Anil K., Robert P. W. DUIN, and Jianchang MAO, 2000. Statistical Pattern Recognition: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(1), 4–37.

JORDAN, Michael I., and Robert A. JACOBS, 1993. Hierarchical Mixtures of Experts and the EM Algorithm. *In*: *IJCNN '93-Nagoya. Proceedings of 1993 International Joint Conference on Neural Networks. Volume 2*. JNNS. pp. 1339–1344.

KITTLER, J., 1998. Combining Classifiers: A Theoretical Framework. *Pattern Analysis and Applications*, **1**(1), 18–27.

KITTLER, J., and F. M. ALKOOT, 2003. Sum versus Vote Fusion in Multiple Classifier Systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**(1), 110–115.

KITTLER, Josef, *et al.*, 1998. On Combining Classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**(3), 226–239.

KLEINBERG, E. M., 1990. Stochastic Discrimination. *Annals of Mathematics and Artificial Intelligence*, **1**(1–4), 207–239.

KLEINBERG, Eugene M., 2000. On the Algorithmic Implementation of Stochastic Discrimination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(5), 473–490.

KROGH, Anders, and Jesper VEDELSBY, 1995. Neural Network Ensembles, Cross Validation, and Active Learning. *In*: Gerald TESAURO, David S. TOURETZKY, and Todd K. LEEN, eds. *Advances in Neural Information Processing Systems 7*. Cambridge, MA: The MIT Press, pp. 231–238.

KUNCHEVA, Ludmila I., 2002a. A Theoretical Study on Six Classifier Fusion Strategies. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(2), 281–286.

KUNCHEVA, Ludmila I., 2002b. Switching Between Selection and Fusion in Combining Classifiers: An Experiment. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, **32**(2), 146–156.

KUNCHEVA, Ludmila I., 2005. Diversity in Multiple Classifier Systems. *Information Fusion*, **6**(1), 3–4.

KUNCHEVA, Ludmila I., and Lakhmi C. JAIN, 2000. Designing Classifier Fusion Systems by Genetic Algorithms. *IEEE Transactions on Evolutionary Computation*, **4**(4), 327–336.

KUNCHEVA, Ludmila I., and Juan J. RODRÍGUEZ, 2007. Classifier Ensembles with a Random Linear Oracle. *IEEE Transactions on Knowledge and Data Engineering*, **19**(4), 500–508.

KUNCHEVA, Ludmila I., and Dmitry P. VETROV, 2006. Evaluation of Stability of k-Means Cluster Ensembles with Respect to Random Initialization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**(11), 1798–1808.

KUNCHEVA, Ludmila I., and Christopher J. WHITAKER, 2003. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Machine Learning*, **51**(2), 181–207.

KUNCHEVA, L. I., *et al.*, 2000. Is Independence Good For Combining Classifiers? *In*: A. SANFELIU, *et al.*, eds. *Proceedings, 15th International Conference on Pattern Recognition, Volume 2*. Los Alamitos: IEEE Computer Society, pp. 168–171.

KUNCHEVA, L. I., *et al.*, 2003. Limits on the Majority Vote Accuracy in Classifier Fusion. *Pattern Analysis and Applications*, **6**(1), 22–31.

LAM, Louisa, and Ching Y. SUEN, 1995. Optimal Combination of Pattern Classifiers. *Pattern Recognition Letters*, **16**(9), 945–954.

LAM, Louisa, and Ching Y. SUEN, 1997. Application of Majority Voting to Pattern Recognition: An Analysis of Its Behavior and Performance. *IEEE Transactions on Systems, Man and Cybernetics, Part A*, **27**(5), 553–568.

LARKEY, Leah S., and W. Bruce Croft, 1997. Combining Classifiers in Text Categorization. *In*: Hans-Peter Frei, *et al.*, eds. *SIGIR '96: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM Press, pp. 289–297.

LEAP, Nathan J., *et al.*, 2008. An Investigation of the Effects of Correlation and Autocorrelation on Classifier Fusion and Optimal Classifier Ensembles. *International Journal of General Systems*, **37**(4), 475–498.

LIU, Y., and X. YAO, 1999. Ensemble Learning via Negative Correlation. *Neural Networks*, **12**(10), 1399–1404.

LIU, Y., X. YAO, and T. HIGUCHI, 2000. Evolutionary Ensembles with Negative Correlation Learning. *IEEE Transactions on Evolutionary Computation*, **4**(4), 380–387.

MELVILLE, Prem, and Raymond J. MOONEY, 2005. Creating Diversity in Ensembles Using Artificial Data. *Information Fusion*, **6**(1), 99–111.

MILLER, David J., and Lian YAN, 1999. Critic-Driven Ensemble Classification. *IEEE Transactions on Signal Processing*, **47**(10), 2833–2844.

OPITZ, David, and Richard MACLIN, 1999. Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence Research*, **11**, 169–198.

OPITZ, David W., 1999. Feature Selection for Ensembles. *In*: American Association for Artificial Intelligence (AAAI), ed. *AAAI-99: Proceedings of the Sixteenth National Conference on Artificial Intelligence*. Menlo Park, CA: AAAI Press, pp. 379–384.

PERRONE, Michael P., and Leon N. COOPER, 1993. When Networks Disagree: Ensemble Methods for Hybrid Neural Networks. *In*: R. J. MAMMONE, ed. *Neural Networks for Speech and Image Processing*. London: Chapman-Hall, pp. 126–142.

RAFTERY, Adrian E., David MADIGAN, and Jennifer A. HOETING, 1997. Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, **92**(437), 179–191.

REYZIN, Lev, and Robert E. SCHAPIRE, 2006. How Boosting the Margin can also Boost Classifier Complexity. *In*: *ICML '06: Proceedings of the 23rd International Conference on Machine Learning*. New York: ACM, pp. 753–760.

RUTA, Dymitr, and Bogdan GABRYS, 2005. Classifier Selection for Majority Voting. *Information Fusion*, **6**(1), 63–81.

SCHAPIRE, Robert E., 1990. The Strength of Weak Learnability. *Machine Learning*, **5**(2), 197–227.

SCHAPIRE, Robert E., 1999. A Brief Introduction to Boosting. *In*: *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, pp. 1401–1406.

SCHAPIRE, Robert E., *et al.*, 1998. Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods. *The Annals of Statistics*, **26**(5), 1651–1686.

SHIPP, Catherine A., and Ludmila I. KUNCHEVA, 2002. Relationships Between Combination Methods and Measures of Diversity in Combining Classifiers. *Information Fusion*, **3**(2), 135–148.

SKURICHINA, Marina, 2001. *Stabilizing Weak Classifiers: Regularization and Combining Techniques in Discriminant Analysis*. Ph. D. thesis, Delft University of Technology, Delft.

SKURICHINA, Marina, and Robert P. W. DUIN, 2002. Bagging, Boosting and the Random Subspace Method for Linear Classifiers. *Pattern Analysis and Applications*, **5**(2), 121–135.

SOLLICH, Peter, and Anders KROGH, 1996. Learning with Ensembles: How Over-Fitting can be Useful. *In*: David S. TOURETZKY, Michael C. MOZER, and Michael E. HASSELMO, eds. *Advances in Neural Information Processing Systems 8*, Bradford Books. Cambridge, MA: The MIT Press, pp. 190–196.

TAX, David M. J., *et al.*, 2000. Combining Multiple Classifiers by Averaging or by Multiplying? *Pattern Recognition*, **33**(9), 1475–1485.

TOPCHY, A., A. K. JAIN, and W. PUNCH, 2003. Combining Multiple Weak Clusterings. *In*: *Third IEEE International Conference on Data Mining, 2003. ICDM 2003.* pp. 331–338.

TOPCHY, Alexander, Anil K. JAIN, and William PUNCH, 2004. A Mixture Model for Clustering Ensembles. *In*: Michael W. BERRY, *et al.*, eds. *Proceedings of the Fourth SIAM International Conference on Data Mining*. Philadelphia, PA: SIAM, pp. 379–390.

TUMER, Kagan, and Joydeep GHOSH, 1996. Analysis of Decision Boundaries in Linearly Combined Neural Classifiers. *Pattern Recognition*, **29**(2), 341–348.

VALENTINI, Giorgio, and Thomas G. DIETTERICH, 2004. Bias-Variance Analysis of Support Vector Machines for the Development of SVM-Based Ensemble Methods. *Journal of Machine Learning Research*, **5**, 725–775.

VALENTINI, Giorgio, and Francesco MASULLI, 2002. Ensembles of Learning Machines. *In*: Maria MARINARO and Roberto TAGLIAFERRI, eds. *Neural Nets: 13th Italian Workshop on Neural Nets, WIRN VIETRI 2002, Vietri sul Mare, Italy, May 30–June 1, 2002. Revised Papers*, Volume 2486 of *Lecture Notes in Computer Science*. Berlin: Springer, pp. 3–19.

WITTNER, Ben S., and John S. DENKER, 1988. Strategies for Teaching Layered Networks Classification Tasks. *In*: Dana Z. ANDERSON, ed. *Neural Information Processing Systems, Denver, Colorado, USA, 1987*. New York: American Institute of Physics, pp. 850–859.

WOLPERT, David H., 1992. Stacked Generalization. *Neural Networks*, **5**(2), 241–259.

WOODS, Kevin, W. Philip KEGELMEYER, Jr, and Kevin BOWYER, 1997. Combination of Multiple Classifiers Using Local Accuracy Estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**(4), 405–410.

XU, Lei, Adam KRZYŻAK, and Ching Y. SUEN, 1992. Methods of Combining Multiple Classifiers and their Applications to Handwriting Recognition. *IEEE Transactions on Systems, Man and Cybernetics*, **22**(3), 418–435.

ZHANG, Chun-Xia, and Jiang-She ZHANG, 2008. A Local Boosting Algorithm for Solving Classification Problems. *Computational Statistics & Data Analysis*, **52**(4), 1928–1941.