UCL DEPARTMENT OF
COMPUTER SCIENCE

Research Note
RN/11/02

Ensemble Learning

20 January 2011

Martin Sewell

Abstract

This note presents a chronological review of the literature on ensemble learning which has accumulated over the past twenty years. The idea of ensemble learning is to employ multiple learners and combine their predictions. If we have a committee of *M* models with uncorrelated errors, simply by averaging them the average error of a model can be reduced by a factor of *M*. Unfortunately, the key assumption that the errors due to the individual models are uncorrelated is unrealistic; in practice, the errors are typically highly correlated, so the reduction in overall error is generally small. However, by making use of Cauchy's inequality, it can be shown that the expected committee error will not exceed the expected error of the constituent models. In this article the literature in general is reviewed, with, where possible, an emphasis on both theory and practical advice, then a taxonomy is provided, and finally four ensemble methods are covered in greater detail: bagging, boosting (including AdaBoost), stacked generalization and the random subspace method.

1 Introduction

The idea of ensemble learning is to employ multiple learners and combine their predictions. If we have a committee of M models with uncorrelated errors, simply by averaging them the average error of a model can be reduced by a factor of M. Unfortunately, the key assumption that the errors due to the individual models are uncorrelated is unrealistic; in practice, the errors are typically highly correlated, so the reduction in overall error is generally small. However, by making use of Cauchy's inequality, it can be shown that the expected committee error will not exceed the expected error of the constituent models. The concept of a committee arises naturally in a Bayesian framework. For a Bayesian—someone who is willing to put a probability on a hypothesis—the task of ensemble learning is, in principle, straightforward: one should use Bayesian model averaging (BMA). This involves taking an average over all models, with each model's prediction weighted by its posterior probability. In BMA it is assumed that a single model generated the whole data set, and the probability distribution over our candidate models simply reflects our uncertainty as to which model that is. By contrast, when we combine multiple models, we are assuming that different data points within the data set can potentially be generated from different models.

2 Literature Review

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

Schapire (1990) introduced *boosting* (the method is explained in Section 5 (page 6)). 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 (explained in Section 6 (page 8)). Xu et al. (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 et al. (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.

In 1995, Yoav Freund and Robert E. Schapire introduced AdaBoost (Freund and Schapire, 1997) (covered in Section 5.1 (page 8)). Cho and Kim (1995) combined the results from multiple neural networks using fuzzy logic which resulted in more accurate classification. Freund (1995) developed a more efficient version of boosting. 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.' Bishop (1995) covers the theoretical aspects of committees of neural networks in general terms, as covered in the Introduction. Krogh and Vedelsby (1995) showed that there is a lot to be gained from using unlabeled data when training ensembles.

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) performed experiments using AdaBoost. Their main conclusion was that boosting performs significantly and uniformly better than bagging when the weak learning algorithm generates fairly simple classifiers. When combined with C4.5, boosting still seemed to outperform bagging slightly, but the results were less compelling. Breiman (1996) introduced bagging (Section 4 (page 6)).

Raftery et al. (1997) consider the problem of accounting for model uncertainty in linear regression models and offer two extensions to BMA: Occam's window and Markov chain Monte Carlo. Woods et al. (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 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 (briefly explained in Section 7 (page 8)). 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) gave an introduction to AdaBoost, and explained 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) considered feature selection for ensembles. Miller and Yan (1999) developed a critic-driven ensemble for classification. Hoeting et al. (1999) wrote a tutorial on BMA. Liu and Yao (1999) presented negative correlation learning for neural network ensembles.

Jain et al. (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, which I have reproduced in 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) considered 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 et al. (2000) presented evolutionary ensembles with negative correlation learning (EENCL). Allwein et al. (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) gave 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. Fumera and Roli (2002) report a theoretical and experimental comparison between two widely used combination rules for classifier fusion: simple average and weighted average of classifiers outputs. They showed that weighted averaging was superior, especially for imbalanced classifiers. 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 et al. (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 better 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 an interesting paper, Evgeniou et al. (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 et al. (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. 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 et al. (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 et al. (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) considered 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. Mease and Wyner (2008) argue experimentally that the statistical view of boosting does not account for its success. They make the counter-intuitive recommendation that if stumps are causing overfitting, be willing to try larger trees. In other words, if boosting a low-complexity base learner leads to overfitting, try a higher-complexity base learner; boosting it might not lead to overfitting. They present empirical evidence to back up their claim. 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. Friedman and Popescu (2008) developed 'rule ensembles'. General regression and classification models are constructed as linear combinations of simple rules derived from the data, and each rule consists of a conjunction of a small number of simple statements concerning the values of individual input variables. These rule ensembles were shown to produce predictive accuracy comparable to the best methods. However, their principal advantage is that they simplify the interpretability of tree ensembles by selecting just a few nodes across all trees. Read et al. (2008) introduced a new method for multi-label classification which uses a pruning procedure to focus on core relationships within multi-label sets. The procedure reduces the complexity and potential for error associated with dealing with a large number of infrequent sets. By combining pruned sets in an ensemble scheme, new label sets can be formed to adapt to irregular or complex data. The method achieved better performance and trained much faster than other multi-label methods.

Domeniconi and Al-Razgan (2009) applied ensembles to clustering, and address the problem of combining multiple weighted clusters that belong to different subspaces of the input space. They leverage the diversity of the input clusterings in order to generate a consensus partition that is superior to the participating ones. Their solutions were as good as or better than the best individual clustering, provided that the input clusterings were diverse. Chen and

Ren (2009) successfully applied bagging to Gaussian process regression models. Ulaş et al. (2009) discuss two approaches to incrementally constructing an ensemble. In an ensemble of classifiers, they choose a subset from a larger set of base classifiers. In an ensemble of discriminants, they choose a subset of base discriminants, where a discriminant output of a base classifier by itself is assessed for inclusion in the ensemble. They found that an incremental ensemble has higher accuracy than bagging and the random subspace method; and it has a comparable accuracy to AdaBoost, but fewer classifiers. Leistner et al. (2009) proposed a novel boosting algorithm, On-line GradientBoost, which outperformed On-line AdaBoost on standard machine learning problems and common computer vision applications. Hido et al. (2009) proposed an ensemble algorithm 'Roughly Balanced (RB) Bagging' using a novel sampling technique to improve the original bagging algorithm for data sets with skewed class distributions. In experiments using benchmark and real-world data sets they compared RB Bagging with the Exactly Balanced Model and other well-known algorithms such as AdaBoost and RIPPER for imbalanced data, and found that RB Bagging generally outperformed them.

Magnus et al. (2010) contrasts BMA with a new method called weighted-average least squares (WALS) with an application to growth empirics. The authors claims that WALS has two major advantages over BMA: its computational burden is trivial and it is based on a transparent definition of prior ignorance. Bühlmann and Hothorn (2010) propose Twin Boosting, which involves a first round of classical boosting followed by a second round of boosting which is forced to resemble the one from the first round. The method has much better feature selection behaviour than boosting, particularly with respect to reducing the number of false positives (falsely selected features). Shalev-Shwartz and Singer (2010) prove that weak learnability is equivalent to linear separability with ℓ_1 margin. This perspective sheds new light on known soft-margin boosting algorithms and enables them to derive several new relaxations of the notion of linear separability. They describe and analyze an efficient boosting framework that can be used for minimizing the loss functions derived from their family of relaxations. In particular, they obtain efficient boosting algorithms for maximizing hard and soft versions of the ℓ_1 margin. Shen and Li (2010) show that the Lagrange dual problems of AdaBoost, LogitBoost and soft-margin LPBoost with generalized hinge loss are all entropy maximization problems. They theoretically and empirically demonstrate that the success of AdaBoost relies on maintaining a better margin distribution. Based on the dual formulation, a general column generation based optimization framework is proposed, which exhibits significantly faster convergence speed than conventional AdaBoost.

3 Taxonomy

There is no definitive taxonomy of ensemble learning. Jain et al. (2000) list eighteen classifier combination schemes; Witten and Frank (2005) detail four methods of combining multiple models: bagging, boosting, stacking and error-correcting output codes; Bishop (2006) covers BMA, committees, boosting, tree-based models and conditional mixture models; Marsland (2009) covers boosting (AdaBoost and stumping), bagging (including subagging) and the mixture of experts method; whilst Alpaydin (2010) covers seven methods of combining multiple learners: voting, error-correcting output codes, bagging, boosting, mixtures of experts, stacked generalization and cascading. The taxonomy in Jain et al. (2000) is repeated in Table 1 (page 7).

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

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

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

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

Table 1: *Ensemble methods (Jain et al., 2000)*

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, 1997), 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.

6 Stacked Generalization

Stacked generalization (or *stacking*) (Wolpert, 1992) is a distinct 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.

8 Conclusion

If you are in a position to put a posterior probability on each of a committee of models, use Bayesian model averaging. Otherwise, the success of ensemble learning relies on the assumption that the errors of the individual models are uncorrelated. In a low noise regime, boosting is to be preferred to bagging, whilst in a high noise regime, bagging is to be preferred to boosting.

References

Allwein, E. L., Schapire, R. E. and Singer, Y. (2000), Reducing multiclass to binary: A unifying approach for margin classifiers, *Journal of Machine Learning Research* **1**, 113–141.

Alpaydin, E. (2010), *Introduction to Machine Learning*, Adaptive Computation and Machine Learning, second edition, The MIT Press, Cambridge, MA.

Battiti, R. and Colla, A. M. (1994), Democracy in neural nets: Voting schemes for classification, *Neural Networks* **7**(4), 691–707.

Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford.

Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, Information Science and Statistics, Springer, New York.

Breiman, L. (1996), Bagging predictors, *Machine Learning* **24**(2), 123–140.

Bühlmann, P. and Hothorn, T. (2007), Boosting algorithms: Regularization, prediction and model fitting, *Statistical Science* **22**(4), 477–505.

Bühlmann, P. and Hothorn, T. (2010), Twin Boosting: Improved feature selection and prediction, *Statistics and Computing* **20**(2), 119–138.

Canuto, A. M. P., Abreu, M. C. C., de Melo Oliveira, L., Xavier, Jr, J. C. and de M. Santos, A. (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, A. and Yao, X. (2006), Evolving hybrid ensembles of learning machines for better generalisation, *Neurocomputing* **69**(7–9), 686–700.

Chawla, N. V., Hall, L. O., Bowyer, K. W. and Kegelmeyer, W. P. (2004), Learning ensembles from bites: A scalable and accurate approach, *Journal of Machine Learning Research* **5**, 421–451.

Chen, T. and Ren, J. (2009), Bagging for Gaussian process regression, *Neurocomputing* **72**(7–9), 1605–1610.

Cho, S.-B. and Kim, J. H. (1995), Multiple network fusion using fuzzy logic, *IEEE Transactions on Neural Networks* **6**(2), 497–501.

Claeskens, G. and Hjort, N. L. (2008), *Model Selection and Model Averaging*, Vol. 27 of *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge University Press, Cambridge.

Dietterich, T. 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, T. G. (2002), Ensemble Learning, *in* M. A. Arbib (ed.), *The handbook of brain theory and neural networks*, second edition, Bradford Books, The MIT Press, Cambridge, MA, pp.405–408.

Domeniconi, C. and Al-Razgan, M. (2009), Weighted cluster ensembles: Methods and analysis, *ACM Transactions on Knowledge Discovery from Data* **2**(4), Article 17.

Džeroski, S. and Ženko, B. (2004), Is combining classifiers with stacking better than selecting the best one?, *Machine Learning* **54**(3), 255–273.

Evgeniou, T., Pontil, M. and Elisseeff, A. (2004), Leave one out error, stability, and generalization of voting combinations of classifiers, *Machine Learning* **55**(1), 71–97.

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

Freund, Y. and Schapire, R. E. (1996), Experiments with a new boosting algorithm, *in* L. Saitta (ed.), *Machine Learning: Proceedings of the Thirteenth International Conference (ICML '96)*, Morgan Kaufmann, San Francisco, CA, pp.148–156.

Freund, Y. and Schapire, R. E. (1997), A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences* **55**(1), 119–139.

Friedman, J. H. and Popescu, B. E. (2008), Predictive learning via rule ensembles, *The Annals of Applied Statistics* **2**(3), 916–954.

Fumera, G. and Roli, F. (2002), Performance analysis and comparison of linear combiners for classifier fusion, *in* T. Caelli, A. Amin, R. P. W. Duin, M. Kamel and D. de Ridder (eds.), *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshops, SSPR 2002 and SPR 2002, Windsor, Ontario, Canada, August 2002, Proceedings*, Vol. 2396 of *Lecture Notes in Computer Science*, Berlin, pp.424–432.

Fumera, G. and Roli, F. (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, N., Hervás-Martnez, C. and Ortiz-Boyer, D. (2005), Cooperative coevolution of artificial neural network ensembles for pattern classification, *IEEE Transactions on Evolutionary Computation* **9**(3), 271–302.

Hadjitodorov, S. T., Kuncheva, L. I. and Todorova, L. P. (2006), Moderate diversity for better cluster ensembles, *Information Fusion* **7**(3), 264–275.

Hansen, B. E. (2007), Least squares model averaging, *Econometrica* **75**(4), 1175–1189.

Hansen, L. K. and Salamon, P. (1990), Neural network ensembles, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12**(10), 993–1001.

Hido, S., Kashima, H. and Takahashi, Y. (2009), Roughly balanced bagging for imbalanced data, *Statistical Analysis and Data Mining* **2**(5–6), 412–426.

Ho, T. K. (1998), The random subspace method for constructing decision forests, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(8), 832–844.

Ho, T. K., Hull, J. J. and Srihari, S. N. (1994), Decision combination in multiple classifier systems, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **16**(1), 66–75.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999), Bayesian model averaging: A tutorial, *Statistical Science* **14**(4), 382–401.

Jain, A. K., Duin, R. P. W. and Mao, J. (2000), Statistical pattern recognition: A review, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(1), 4–37.

Jordan, M. I. and Jacobs, R. A. (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 Alkoot, F. M. (2003), Sum versus vote fusion in multiple classifier systems, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**(1), 110–115.

Kittler, J., Hatef, M., Duin, R. P. W. and Matas, J. (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, E. M. (2000), On the algorithmic implementation of stochastic discrimination, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(5), 473–490.

Krogh, A. and Vedelsby, J. (1995), Neural network ensembles, cross validation, and active learning, *in* G. Tesauro, D. S. Touretzky and T. K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, The MIT Press, Cambridge, MA, pp.231–238.

Kuncheva, L. I. (2002a), A theoretical study on six classifier fusion strategies, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(2), 281–286.

Kuncheva, L. 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, L. I. (2005), Diversity in multiple classifier systems, *Information Fusion* **6**(1), 3–4.

Kuncheva, L. I. and Jain, L. C. (2000), Designing classifier fusion systems by genetic algorithms, *IEEE Transactions on Evolutionary Computation* **4**(4), 327–336.

Kuncheva, L. I. and Rodríguez, J. J. (2007), Classifier ensembles with a random linear oracle, *IEEE Transactions on Knowledge and Data Engineering* **19**(4), 500–508.

Kuncheva, L. I. and Vetrov, D. P. (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, L. I. and Whitaker, C. J. (2003), Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, *Machine Learning* **51**(2), 181–207.

Kuncheva, L. I., Whitaker, C. J., Shipp, C. A. and Duin, R. P. W. (2000), Is independence good for combining classifiers?, *in* A. Sanfeliu, J. J. Villanueva, M. Vanrell, R. Alquezar, A. K. Jain and J. Kittler (eds.), *Proceedings, 15th International Conference on Pattern Recognition, Volume 2*, IEEE Computer Society, Los Alamitos, pp.168–171.

Kuncheva, L. I., Whitaker, C. J., Shipp, C. A. and Duin, R. P. W. (2003), Limits on the majority vote accuracy in classifier fusion, *Pattern Analysis and Applications* **6**(1), 22–31.

Lam, L. and Suen, C. Y. (1995), Optimal combination of pattern classifiers, *Pattern Recognition Letters* **16**(9), 945–954.

Lam, L. and Suen, C. Y. (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, L. S. and Croft, W. B. (1997), Combining classifiers in text categorization, *in* H.-P. Frei, D. Harman, P. Schäuble and R. Wilkinson (eds.), *SIGIR '96: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, New York, pp.289–297.

Leap, N. J., Clemans, P. P., Bauer, Jr, K. W. and Oxley, M. E. (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.

Leistner, C., Saffari, A., Roth, P. M. and Bischof, H. (2009), On robustness of on-line boosting - a competitive study, *in* T. Gevers, C. Rother, S. Tominaga, J. van de Weijer and T. Zickler (eds.), *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, IEEE, Piscataway, NJ, pp.1362–1369.

Liu, Y. and Yao, X. (1999), Ensemble learning via negative correlation, *Neural Networks* **12**(10), 1399–1404.

Liu, Y., Yao, X. and Higuchi, T. (2000), Evolutionary ensembles with negative correlation learning, *IEEE Transactions on Evolutionary Computation* **4**(4), 380–387.

Magnus, J. R., Powell, O. and Prüfer, P. (2010), A comparison of two model averaging techniques with an application to growth empirics, *Journal of Econometrics* **154**(2), 139–153.

Marsland, S. (2009), *Machine Learning: An Algorithmic Perspective*, Chapman & Hall/CRC Machine Learning & Pattern Recognition, CRC Press, Boca Raton.

Mease, D. and Wyner, A. (2008), Evidence contrary to the statistical view of boosting, *Journal of Machine Learning Research* **9**, 131–156.

Melville, P. and Mooney, R. J. (2005), Creating diversity in ensembles using artificial data, *Information Fusion* **6**(1), 99–111.

Miller, D. J. and Yan, L. (1999), Critic-driven ensemble classification, *IEEE Transactions on Signal Processing* **47**(10), 2833–2844.

Opitz, D. and Maclin, R. (1999), Popular ensemble methods: An empirical study, *Journal of Artificial Intelligence Research* **11**, 169–198.

Opitz, D. 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*, AAAI Press, Menlo Park, CA, pp.379–384.

Perrone, M. P. and Cooper, L. N. (1993), When networks disagree: Ensemble methods for hybrid neural networks, *in* R. J. Mammone (ed.), *Neural Networks for Speech and Image Processing*, Chapman-Hall, London, pp.126–142.

Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997), Bayesian model averaging for linear regression models, *Journal of the American Statistical Association* **92**(437), 179–191.

Read, J., Pfahringer, B. and Holmes, G. (2008), Multi-label classification using ensembles of pruned sets, *in* F. Giannotti, D. Gunopulos, F. Turini, C. Zaniolo, N. Ramakrishnan and X. Wu (eds.), *Proceedings. Eighth IEEE International Conference on Data Mining. ICDM 2008*, IEEE, Los Alamitos, pp.995–1000.

Reyzin, L. and Schapire, R. E. (2006), How boosting the margin can also boost classifier complexity, *in ICML '06: Proceedings of the 23rd International Conference on Machine Learning*, ACM, New York, pp.753–760.

Ruta, D. and Gabrys, B. (2005), Classifier selection for majority voting, *Information Fusion* **6**(1), 63–81.

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

Schapire, R. E. (1999), A brief introduction to boosting, *in Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, pp.1401–1406.

Schapire, R. E., Freund, Y., Bartlett, P. and Lee, W. S. (1998), Boosting the margin: A new explanation for the effectiveness of voting methods, *The Annals of Statistics* **26**(5), 1651–1686.

Shalev-Shwartz, S. and Singer, Y. (2010), On the equivalence of weak learnability and linear separability: New relaxations and efficient boosting algorithms, *Machine Learning* **80**(2–3), 141–163.

Shen, C. and Li, H. (2010), On the dual formulation of boosting algorithms, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32**(12), 2216–2231.

Shipp, C. A. and Kuncheva, L. I. (2002), Relationships between combination methods and measures of diversity in combining classifiers, *Information Fusion* **3**(2), 135–148.

Skurichina, M. (2001), Stabilizing Weak Classifiers: Regularization and Combining Techniques in Discriminant Analysis, PhD thesis, Delft University of Technology, Delft.

Skurichina, M. and Duin, R. P. W. (2002), Bagging, boosting and the random subspace method for linear classifiers, *Pattern Analysis and Applications* **5**(2), 121–135.

Sollich, P. and Krogh, A. (1996), Learning with ensembles: How over-fitting can be useful, *in* D. S. Touretzky, M. C. Mozer and M. E. Hasselmo (eds.), *Advances in Neural Information Processing Systems 8*, Bradford Books, The MIT Press, Cambridge, MA, pp.190–196.

Tax, D. M. J., van Breukelen, M., Duin, R. P. W. and Kittler, J. (2000), Combining multiple classifiers by averaging or by multiplying?, *Pattern Recognition* **33**(9), 1475–1485.

Topchy, A., Jain, A. K. and Punch, W. (2003), Combining multiple weak clusterings, *in Third IEEE International Conference on Data Mining, 2003. ICDM 2003.*, pp.331–338.

Topchy, A., Jain, A. K. and Punch, W. (2004), A mixture model for clustering ensembles, *in* M. W. Berry, U. Dayal, C. Kamath and D. Skillicorn (eds.), *Proceedings of the Fourth SIAM International Conference on Data Mining*, SIAM, Philadelphia, PA, pp.379–390.

Tumer, K. and Ghosh, J. (1996), Analysis of decision boundaries in linearly combined neural classifiers, *Pattern Recognition* **29**(2), 341–348.

Ulaş, A., Semerci, M., Yildiz, O. T. and Alpaydin, E. (2009), Incremental construction of classifier and discriminant ensembles, *Information Sciences* **179**(9), 1298–1318.

Valentini, G. and Dietterich, T. G. (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, G. and Masulli, F. (2002), Ensembles of learning machines, *in* M. Marinaro and R. Tagliaferri (eds.), *Neural Nets: 13th Italian Workshop on Neural Nets, WIRN VIETRI 2002, Vietri sul Mare, Italy, May 30–June 1, 2002. Revised Papers*, Vol. 2486 of *Lecture Notes in Computer Science*, Springer, Berlin, pp.3–19.

Witten, I. H. and Frank, E. (2005), *Data Mining: Practical Machine Learning Tools and Techniques*, The Morgan Kaufmann Series in Data Management Systems, second edition, Morgan Kaufmann, San Francisco, CA.

Wittner, B. S. and Denker, J. S. (1988), Strategies for teaching layered networks classification tasks, *in* D. Z. Anderson (ed.), *Neural Information Processing Systems, Denver, Colorado, USA, 1987*, American Institute of Physics, New York, pp.850–859.

Wolpert, D. H. (1992), Stacked generalization, *Neural Networks* **5**(2), 241–259.

Woods, K., Kegelmeyer, Jr, W. P. and Bowyer, K. (1997), Combination of multiple classifiers using local accuracy estimates, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19**(4), 405–410.

Xu, L., Krzyżak, A. and Suen, C. Y. (1992), Methods of combining multiple classifiers and their applications to handwriting recognition, *IEEE Transactions on Systems, Man and Cybernetics* **22**(3), 418–435.

Zhang, C.-X. and Zhang, J.-S. (2008), A local boosting algorithm for solving classification problems, *Computational Statistics & Data Analysis* **52**(4), 1928–1941.