Multiple Classifier Combination: Lessons and Next Steps *

Tin Kam Ho
Bell Laboratories, Lucent Technologies
700 Mountain Avenue, 2C425, Murray Hill, NJ 07974, USA
Email: tkh@bell-labs.com

August 31, 2001

Abstract

During the 1990's many methods were proposed for combining multiple classifiers for a single recognition task. With these methods, the focus of the field shifted from the competition among specific statistical, syntactic, or structural approaches to the integration of all these as potential contributing components in a combined system. Deeper explorations of the combination methods revealed many links back to several fundamental issues of pattern recognition. Amid the excitement and confusion, there is a persistent uncertainty concerning the optimal match between a method and a problem, due to a strong dependence of classifier performance on the data. In this chapter I review several different motivations that have driven this development, summarize lessons learned in the exploration of combination methods, outline the difficulties encountered, and suggest ways to break out of the current plateau.

1 Introduction

During the past decade the method of multiple classifier systems was firmly established as a practical and effective solution for difficult pattern recognition tasks. The idea appeared under many names: hybrid methods, decision combination, multiple experts, mixture of experts, classifier ensembles, cooperative agents, opinion pool, sensor fusion, and more. In some areas it was motivated by an empirical observation that specialized classifiers often excel in different cases and make different errors. In other areas it arose naturally from the application context, such as the need to employ a variety of sensor types, which induces a natural decomposition of the problem. In these cases, the input can be considered as vectors projected to the subspaces of the full feature space, which have to be matched by different procedures. An example is authentication by fingerprints together with voice, gestures etc., or retrieving a document by string matching on keywords and by matching word count vectors. There were also cases where the motivation was to avoid having to make a meaningful choice of some arbitrary initial condition, such as the initial weights for a neural network. Or, disturbingly, there were arguments that introducing some randomness, no matter where, in classifier training is a good thing to do, just to make a collection containing some differences that, when combined, will automatically perform better than a single element.

There are many ways to use more than one classifier in a single recognition problem. A divide-and-conquer approach would isolate the types of input for which each classifier performs well, and direct new inputs of a particular type to the suitable classifier. A sequential approach would use one classifier first, and invoke others only if it fails to yield a decision with sufficient

*To appear in A. Kandel, H. Bunke, (eds.), *Hybrid Methods in Pattern Recognition*, World Scientific, 2002

confidence. All these can be said to be multiple classifier strategies, and have been explored to a certain extent. However, motivated by the factors mentioned earlier, the majority of classifier combination research focuses on applying all the available classifiers in parallel to the same input and combining their decisions. Naturally one asks, what is gained and lost in a parallel combination? When is it preferable to those alternative approaches?

As the idea prospered, many different combination techniques emerged. It almost feels like we are simply bringing the fundamental pursuit to a different level. Instead of looking for the best set of features and the best classifier, now we look for the best set of classifiers and then the best combination method. One can imagine that very soon we will be looking for the best set of combination methods and then the best way to use them all. If we do not take the chance to review the fundamental problems arising from this challenge, we are bound to be driven into such an infinite recurrence, dragging along more and more complicated combination schemes and theories, and gradually losing sight of the original problem.

The trend of parallel combination of many classifiers deviates from, or even follows an opposite philosophy of, the traditional selection approach, where one evaluates the available classifiers against a representative sample, and chooses the best one to use. Here, in essence, one abandons the attempt to optimize an individual classifier, and instead, tries to use all the available ones in an intelligent way. This is the antithesis of economical design. Introducing needless classifiers would harm more than just efficiency. The agreement of two poor classifiers does not necessarily yield more correct decisions. And very soon one can fall into a situation that the same training data are used to estimate an increasing and potentially infinite number of classifier parameters, which is not an unfamiliar trap.

So, is classifier combination a well justified, systematic methodology, or is it a desperate attempt to make the most out of imperfect designs? What has been learned in the past decade, what is still missing, and what should we do next?

In this chapter, I review the development of the relevant ideas, discuss the lessons and implications, and try to outline some potentially fruitful future directions. After a decade of development, the literature on classifier combination has become huge, especially as one recognizes in retrospect similar attempts under different names in different communities. The review in this paper is not intended to be comprehensive. Rather, I attempt to point to works exemplifying several main ideas, with emphasis on distinguishing between methods appearing under apparently similar names, and pointing out their deficiencies and difficulties. For a more complete coverage, especially on successful, empirical examples, readers are referred to several recent survey articles, proceedings, and books [14] [43] [48] [54] [69].

I will start with the historical context most familiar to researchers and practitioners in pattern recognition, and extend into similar attempts in related disciplines. I will focus on the methods first and analyses later, since in most cases discoveries of the methods occurred first, and theoretical analyses followed only after much empirical exploration. There were notable exceptions to this where theoretical and algorithmic development were tightly coupled, for which the two lines of discussion cannot be separated.

2 The Historical Context: Awareness of the Advantages of Multiple Approaches

By the end of the 1980's, the field of pattern recognition had developed into a rich area featuring several active branches. There were the syntactic methods integrating results from formal languages and automata theory. There were the structural approaches focusing on innovative feature representations and matching procedures. Statistical approaches continued to provide a core depository of methods, and were greatly supplemented by vivid activities using the neural network paradigm or following the machine learning themes.

For a practitioner facing the diverse possibilities, selecting among these methods became a major challenge. In the absence of a consistent and dominant winner, each method had to be evaluated in the context of each new problem, within the constraints of many practical concerns. Efforts in designing a recognition system emphasized selecting the best classifier for

a given task, supplemented occasionally by divide-and-conquer strategies that invoke different classifiers for different inputs under guidance of considerations such as confidence of decisions, cost of application, or other specific knowledge on the input. An early example is a two-classifier system that invokes a nearest neighbor classifier only when a linear classifier decides that the input falls in an ambiguous region [11].

The complementary advantages of different classification methods were recognized early on. Aside from efficiency concerns, the potential of using different feature representation and decision rules for the same problem was appealing. A review by Bunke [9] cited many of the early efforts in incorporating elements of one approach into another, such as combining statistical and syntactic methods through probabilistic and attributed grammars [23], matching structural prototypes by syntactic methods via graph grammars [27], or treating structural matching as a constraint satisfaction process. Influenced by popular ideas from artificial intelligence research, recognition was also seen as a process of inferences about class concepts from concrete examples. The inference steps can be organized as systems of production rules, which allows flexibility in including multiple sources of knowledge and multiple types of evidence. There, the emphasis was on knowledge representation methods and control procedures for the inference steps. Some attempted to model recognition as constraint satisfaction processes. The awareness of the utility of diverse knowledge sources also led to studies on fusing sensory data of multiple types [58] [64], or combining the opinions of multiple experts using consensus theory [3].

The idea of mixing different classifiers for the same problem continued to evolve into hierarchical or multi-stage systems, where each stage identifies or rejects a subset of classes, or more sophisticated conditional systems where different classifiers are selectively applied or believed according to different characteristics of the input [82] or agreement on decisions [34]. In mixture of experts [62] a gating network was added to select among the outputs of a number of expert networks each of which specializes in particular classes or subclasses by way of unsupervised learning. Whereas each expert network learns the association between an input vector and its correct class, a gating network learns the association between the correctness of the decision of those expert networks and features of the input.

Though a number of such proposals made their way into practical systems, most remained interesting possibilities due to a lack of demonstrated, consistent success. Design of each hybrid system was highly specific to the particular application.

3 Reaching a Consensus: Optimization of Combined Decisions

Around the beginning of the 1990's, a more ambitious question emerged: if no classifiers are known to be best for all cases, is it possible to take advantage of the strengths of each method while avoiding their weaknesses? Different from most previous attempts, here the focus is on *simultaneous* applications of multiple classifiers to the same input, and the objective is to improve overall accuracy. In the early days, whether this can be achieved was by no means obvious. Why would one dilute the strengths of the best classifier with the often poor decisions of the inferior ones? How can one be sure that the correct decisions of the stronger classifiers are preserved when the weaker classifiers join in? Simple approaches such as the majority vote, though known to increase decision reliability, do not necessarily give a higher *overall* correct rate; if the system rejects a decision for cases lacking agreement by majority, the overall rate of successful recognition is actually *lower* than that of the best individual.

Accuracy improvement becomes possible only when more information is taken from the classifiers. Instead of just a single decision on one of the N classes assumed for the problem, one needs to consider the secondary choices, or have the classifiers score each class with a degree of belief for the respective input.

Classifiers operating by different principles do not necessarily produce belief scores on a compatible scale, which presents a problem in decision combination. For example, an attempt to bring syntactic and statistical methods into the same system [24] had to rely on an unnatural sum of the incommensurable syntactic distances between the symbol strings and the semantic

distance between the associated parameter vectors. A more justified approach was adopted in evidence combination [59], where statistical models and the Bayes rule were used to transform the distance measures computed by the individual classifiers into belief scores for each class, and then Dempster-Shafer theory of evidence [68] was used to combine the scores. Explorations in applying the Bayesian formalism to combine belief scores continued, making use of detailed knowledge of the classifiers' behavior contained in confusion matrices [83], or accuracies associated with every pattern of joint decisions by all classifiers in the collection [41] [80]. Methods relying on such detailed knowledge of the classifiers' behavior run into difficulties when the number of classes is large, as the number of training samples required to obtain reliable probabilistic estimates becomes prohibitive. A different strategy is to convert the class decisions to a uniform linear scale by reducing all similarity scores to rank orderings of the classes, trading off some precision for robustness. Borda count or logistic regression were used to combine such rank scores, and the latter also helped to identify redundant classifiers in the collection [34].

The problem appears simpler in the context where all classifiers are of the same type. If each classifier computes an estimate of the posterior probability of each class using each feature vector, such estimates can be directly combined by trainable linear or logarithmic functions called the *opinion pool* [2], or by fixed maximum, median, mean, sum, or product rules [47]. Classifiers that decide by principles of hypothesis testing can be combined by methods of meta-analysis [32] or logistic regression [42]. At the extreme, the decisions of individual classifiers, in whatever form, can be considered as inputs to a second level classifier [57].

The idea of employing multiple experts specialized for different aspects of a given task is probably as old as the history of human society. But such common wisdom does not necessarily and immediately apply to the context of pattern recognition, where concepts such as disagreement and cooperation take on a concrete meaning. The behavior of the individual classifiers can be mathematically characterized, and with accuracy being an objective measure of effectiveness, the benefits of any proposal can potentially be quantified. Reasons for having multiple sources of knowledge about an input pattern can be various, but whether they should be maintained in separate representations is never obvious. Even less clear is whether one should integrate the separate representations and compare them under a single metric, or direct them to separate classifiers and defer the integration until after the classifiers have processed them. Regardless of the level where the integration is carried out, details of the integration procedure have to be stated in terms of a concise mathematical function and implemented in a well-defined algorithm. And there are vast differences in the effectiveness of each procedure as observed in the early explorations.

From the engineering perspective, the discovery that the combined accuracy of several classifiers applied simultaneously can be made better than each of the individuals came as a surprise from experiments in the late 1980's and early 1990's. Numerous studies followed on, confirming former observations or leading to new proposals, and by now combining classifiers has become almost a routine. Still, accuracy improvement is not guaranteed, and not all combination schemes are successful. Have we fully understood the successes and failures and all the tradeoffs? In the following I outline some of the attempted analyses.

An early result on committee votes is Condorcet's Jury Theorem in the 18th century [10] [55] which asserts that, if each individual member of the committee has an independent opinion about the subject to be voted on, the probability of the majority vote being correct increases or decreases monotonically with the size of the committee, and converges to one or zero depending on whether each member has a chance of being correct that is greater or less than 50%. Engvall [18] gave a least upper bound for the average accuracy of decisions agreed to by multiple classifiers to a given degree. Srihari [72] [73] related the reliability of the committee decision to the average individual reliabilities and individual differences, and gave the number of voters yielding an optimal combined reliability for some specific cases. These are purely probabilistic arguments. Mazurov et al. [60] gave a geometrical analysis where the voters and their decisions are modeled as a system of linear inequalities. Conditions on the consistence of the inequalities were derived for the existence of a minimal committee achieving a particular fraction of agreed decisions, and methods were given for constructing such minimal committees. Lam and Suen [55] studied expected voting outcomes under different conditions on the committee.

Simple voting schemes such as the majority or plurality rules are fixed regardless of the performance of the voters and the conditions of the input. So these studies focused on evaluation of such rules under different patterns of voter performance. The results show, under a specific performance goal, whether one should use a particular rule or some specific number of voters, but the rules themselves are not trainable.

Votes on non-binary choices can be given as a ranking of the choices (classes in the pattern recognition context). In behavior sciences, methods for combining such rankings are referred to as *social choice functions* or *group consensus functions*. Arrow's impossibility theorem [1] generalized an observation by Condorcet on cyclic majorities occurring when there are transitive preferences on the choices among the individual voters, and asserted that there is no decision combination function for preferences among three or more items by two or more individuals that can satisfy all four axioms capturing the intuition of a reasonable combination. Goodman and Markowitz [29] suggested that by modifying some of the conditions imposed by Arrow, some acceptable combination functions can be obtained. Fishburn reviewed these in detail [19] [20] and extended the theory to infinite sets of individuals. Related mathematical theories were also covered by Black [5]. To quantify the differences between individual decision makers, Bogart [6] [7] presented a theory of distances between partial orders.

In the social scientific theories of votings and elections, the focus is on obtaining a combined choice that best represents the voters' preferences. There is no notion of absolute *correctness* of the combined choice. The merit of a candidate in an election is solely determined by some specific characterization of voters' preferences. However, in classification, there is a true class associated with each input that is determined regardless of the combination mechanism. That makes a difference, since the combination function can be trained to optimize some objective accuracy measure. In classification, the prior performance of the voters can also be evaluated against the objective truth, and based on such evaluation the combination function can be tuned. These observations motivated the use of regression functions to combine the rank decisions [34], where merits of individual preferences are evaluated by their contribution to the correctness of decisions in the past. Degenerate forms of such tuned combination functions became various weighted voting schemes that can be applied to binary choices.

In the context of pattern recognition, the individual decisions of the voters are never independent. They are intrinsically linked by the fact that they are responses to the same input pattern. The degree of agreement of the individual decisions on the same input case, besides characterizing the amount of differences among the classifiers, is also a reflection of the relative difficulty of the case. With a well designed set of classifiers, for an easy case, all or most of the decisions should be the same. Therefore, the pattern of agreement can be used to differentiate the types of input cases [34] [35]. Tubbs and Alltop [75] derived measures of confidence associated with combining multiple ranked lists. More recently, Tumer and Ghosh [77] quantified the improvement of classifier accuracy by biases in the ranked lists. Other inferences about rankings were pursued in the area of order statistics [33] [61].

For combining estimates of posterior probabilities or belief scores represented in a normalized, continuous scale, Bayes decision theory and Dempster-Shafer's theory of evidence [68] dominated. Prior performances of individual classifiers can be embedded into the combination function in the form of accuracy estimates conditioned on the individual decisions. Simpler combination rules that do not take into account the classifiers' prior performance were studied by Kittler et al. [47], where a justification was given for the sum rule that chooses the class maximizing the sum of individual estimates of posterior probabilities. The justification is from the sum rule's relative insensitivity to local estimation errors when compared to the product rule. There are also other methods that essentially treat the belief scores given by the individuals as input features for a classifier at another level. These derive support from the basic principles of classification. In statistics they are known as model-mix methods [56], and are justified by a reduction of the bias of the combined estimator.

The combination methods summarized above all assume that the classifiers are given and unchangeable, or consider them already optimized for the task. The design of the combination function occurs after the classifiers are chosen and their training is completed. The hope is to employ a combination function that maps the individual decisions to a final decision which,

when evaluated over all inputs, has a better chance of being correct. Thus I refer to these as *decision optimization* methods.

A system using several classifiers may not be able to achieve the highest accuracy for a problem if there are cases for which none of the classifiers' decisions is sufficiently close to being correct. A question then arose: is it possible to systematically create a collection of classifiers for a given problem, so that for each possible input pattern there exists at least one member that can correctly identify its true class? Such a classifier should dominate the final decision of the combination to guarantee its effect. If this could be done, accuracy can be maximized for any recognition problem.

This much more ambitious goal, together with other relatively modest hopes of getting more uniform coverage of the input cases and overcoming certain biases in the training process, motivated the development of another family of techniques that I refer to as *coverage optimization* methods. There, instead of the classifiers, the *decision combination function* is assumed to be fixed and unchangeable in form. The strategy is to *create* a set of classifiers, observing some specific notion of complementariness, such that they can yield a good final decision under the chosen combination function.

4 Forcing Complementariness: Generative Methods and Stochastic Searches

Some classifiers have an inherent stochastic component, or places where arbitrary decisions have to be made. Neural network methods, having many architectural and algorithmic possibilities, but less well-understood effects, run into such difficulties more often than other techniques. Many neural networks need initial weights before training begins. Choosing one set of random values means committing to a specific starting point which, in bad cases, could lead to a local optimum. The sequential order in which training samples are presented leads to another line of possibilities. And there are even more subtle decisions like the type of architecture, the number of hidden layers, or the number of units in each layer. The performances of individual networks trained for the same problem can differ by a large amount due to such design choices, and it was observed that they can excel in different areas of the input space. Such observations led to the idea of combining the network decisions by majority or plurality votes [31] [63], where the hope is that the overlap of the correct decisions would dominate the coincidence of errors.

Stacked generalization [81] introduced another possibility. The role of a classifier is understood to be generalizing class concepts from known examples to the unknown. Assuming each particular generalizer carries a certain bias, the method attempts to deduce such biases by way of cross-validation, a well established statistical procedure that, at each pass, reserves random parts of the training data for model evaluation. Such biases are then modeled and reduced by a second level classifier trained on the raw decisions. The idea of learning the classifier bias from within the training set was carried further in the method of bootstrap aggregation, or bagging [8], where each classifier is trained by a bootstrap replicate of the training sample, i.e., sampling the training set to the same size with replacement. The positive effect of bagging is considered to be a result of variance reduction.

The pursuit of complementary errors took an explicit form in the method of boosting [21], a sequential procedure where later classifiers are trained by a subsample of the training set, biased towards the errors committed by the collection of earlier classifiers. Later [44] [71] similar ideas were tried with linear discriminators. Explanations attempted for the effect of the method include distribution of margins and error bounds derived from the VC dimension theory [67], and fitting an additive model by forward stepwise optimization on a criterion similar to binomial log-likelihood [22]. However, as we will see shortly, these attempted explanations are incomplete or almost irrelevant, as they fail to address, simultaneously, all the important factors involved in the process: the discriminating power and generalization ability of individual classifiers, and the correlation among them.

Despite observed successes in many practical experiments, such training set subsampling methods suffer from a logical paradox. Weakening the individual classifiers by not training

on or equally weighing in all available data is said to help avoid overfitting. Boosting simply cannot run on classifiers perfect for the training data, as there are no errors to train additional components. But the design is intended to make the entire system work well on the full training set. So do we want the classifiers perfectly adapt to the training set or not? If we do, what is the point of deferring this adaptation to the level of decision combination? If not, what is the point of adding more and more components to improve accuracy on the full training set? Why should the full system treat the training set differently from the component classifiers? If the training set is assumed to faithfully represent the unseen cases, why would one believe that by sacrificing training set accuracy one can gain testing set accuracy? If the feature space is small and the training sample is dense, the training set could overlap with the testing set perfectly. In such a case, what good will it do to deliberately sacrifice accuracy on the training set? On the other hand, without involving the generalization performance in the analysis, the argument that the methods can, eventually, do perfectly on the training set is useless; template matching can do the job, so there is no reason to bother with such elaborate training procedures and sacrifices. Without a thorough understanding of how overfitting is avoided or controlled within the training process, there is no guarantee on the results, and empirical evidence does show that these methods do not always work.

Then there is the question of the form of the component classifiers. All these methods are known to work well with decision trees, though the specific way data are split at each internal node matters. The much used notion of weak classifiers is not well defined. Fully split decision trees are very strong classifiers, though pruned or forced shallow versions have been used with some success. With others, like linear discriminators, things are less clear. If the component classifiers are too weak, given the simple decision combination function of weighted or unweighted averaging, the decisions of many bad classifiers can easily outweigh the good ones, especially in methods like boosting that focus on early errors. And how about mixing in different types of classifiers?

Such fundamental issues are in the midst of confusion in several communities. Nevertheless, a rigorous analysis of these issues has been given in Kleinberg's theory and method of stochastic discrimination [49] [50], which historically preceded most of the above developments. The analysis uses a set theoretic abstraction to remove all the algorithmic details of classifiers, features, and training procedures. It considers only the classifiers' decision regions in the form of point sets, called weak models, in the feature space. A collection of classifiers is thus just a sample from the power set of the feature space. If the sample satisfies a uniformity condition, i.e., if its coverage is unbiased for any local region of the feature space, then a symmetry is observed between two probabilities (w.r.t. the feature space and w.r.t. the power set, respectively) of the same event that a point of a particular class is covered by a component of the sample. Discrimination between classes is achieved by requiring some minimum difference in each component's inclusion of points of different classes, which is trivial to satisfy. The symmetry translates such differences across different points in the space to differences among the models on a single point. Accuracy in classification is then governed by the law of large numbers. If the sample of weak models is large, the discriminant function, defined on the coverage of the models on a single point and the class-specific differences within each model, converges to poles distinct by class with diminishing variance. Moreover, it is proved that the combined system maintains the projectability of the single weak models, i.e., if each model is thick enough with respect to the spatial continuity of the classes, so that estimates of the point inclusion probabilities from the training set are close to the true probabilities, then the combined system would retain the same goodness of the estimate, which translates, by the symmetry, to accuracy in classifying unseen points. Berlind [4] analyzed an alternative discriminant for multi-class problems.

The theory of stochastic discrimination is complete. It identifies three and only three sufficient conditions for a classifier to achieve maximum accuracy for a problem. Each of the conditions is spelled out in concrete definitions and can be measured in precise terms. Nothing more is needed to reach optimal accuracy. What is good about building the classifier on weak models instead of strong models? Because weak models are easier to obtain, and their smaller capacity renders them less sensitive to sampling errors in small training sets [78] [79]. Why are many models needed? Because the method relies on the law of large numbers to reduce the

variance of the discriminant on each single point. The uniformity condition specifies exactly what kind of correlation is needed among the individual models. Moreover, accuracy is not achieved by intentionally limiting the VC dimension of the complete system; the combination of many weak models can have a very large VC dimension. It is a consequence of the symmetry relating probabilities in the two spaces, and the law of large numbers. It is a structural property of the topology.

The theory includes explicitly each of the three elements long believed to be important in pattern recognition: discrimination power, complementary information, and generalization ability. Here the projectability of the weak models is a key element in the proof, and not an implicit side effect. Methods like boosting, when put in this framework, can be understood as heuristics for improving (but not guaranteeing) the uniformity of model coverage. The lack of an explicit treatment of projectability leads to observed overfitting in some experiments. Bootstrapping just focuses on reducing the variances due to small training samples, and as the training sample gets larger, its usefulness diminishes. Finally, the notion of uniform coverage in stochastic discrimination is rigorously defined, as opposed to many other vague concepts such as diversity, independence, and correlation. Without arguments like these addressing all the three aspects, other schemes for introducing randomness into a collection are, at most, just counting on a blind hope.

As a constructive procedure, the method of stochastic discrimination depends on a detailed control of the uniformity of model coverage, which is outlined but not fully published in the literature [51]. The method of random subspaces followed these ideas but attempted a different approach. Instead of obtaining weak discrimination and projectability through simplicity of the model form, and forcing uniformity by sophisticated algorithms, the method uses complete, locally pure partitions as given in fully split decision trees [36] or nearest neighbor classifiers [37] to achieve strong discrimination and uniformity, and then explicitly forces different generalization patterns on the component classifiers. This is done by training large capacity component classifiers such as nearest neighbors and decision trees to fully fit the data, but restricting the training of each classifier to a coordinate subspace of the feature space, such that classifications remain invariant in the complement subspace. If there is no ambiguity in the subspaces, the individual classifiers maintain maximum accuracy on the training data, with no cases deliberately chosen to be sacrificed, and thus the method does not run into the training set paradox. However the tension among the three factors persists. There is another difficult tradeoff in how much discriminating power to retain for the component classifiers. Can every one use only a single feature dimension so as to maximize invariance in the complement dimensions? Also, projection to coordinate subspaces sets parts of the decision boundaries parallel to the coordinate axes. Augmenting the raw features by simple transformations [36] introduces more flexibility, but it may still be insufficient for an arbitrary problem. Optimization of generalization performance will continue to depend on a detailed control of the projections to suit a particular problem.

Another interesting idea explored was building different classifiers by partitioning the set of classes in different ways [15]. This works directly only for a large number of classes, but the idea may be extensible to subclasses within a smaller number of classes. The method trains each component classifier only to parts of the desired decision boundary, thereby maintaining some generalization power in the unadapted regions. However, it is unclear whether or how discrimination between all pairs of classes can be fully recovered in the combination.

There were also attempts to mix and match all these different ways to generate component classifiers. This, like a few other ideas bearing similar names to the above mentioned, is not backed by any serious theories and does not have a clear motivation. On the other hand, the complicated tradeoffs involved in those better motivated techniques are still open for investigation. The results obtained using decision trees as component classifiers do not always generalize to other types of classifiers, for example, linear discriminators. And what is the exact role of randomization? Can these methods use deterministic algorithms in selecting or omitting specific training cases or feature dimensions? In these methods, the combination of decisions is usually through simple or weighted averaging of the individual probability estimates. Can these methods work with more sophisticated decision combination mechanisms like those explored for given, fixed classifiers?

5 The Lessons

The possibility, by now well supported by empirical evidence, of being able to go beyond the power of traditional classifiers is exciting. But more questions are raised with the proposal of each new method and associated experimentation. Theoretical explanations are either logically incomplete or have to stop at a level where the combinatorics defy detailed modeling. Empirical evidence is often very specific to particular applications, and is seldom systematically documented. So, after more than a decade, where do we stand?

On the choices of methods, we have learned that there are two approaches, decision optimization or coverage optimization. The decision combination methods are at a higher level, as they can be applied to classifiers obtained via the coverage optimization methods. Choices of decision combination methods are dictated by several factors: what type of output the classifiers can produce, how many classes there are, and whether sufficient training data are available to tune the combination function. Table 1 summarizes the best known decision combination methods under joint consideration of two factors.

Table 1: Contextual requirement of various decision combination methods

	Resolution of belief scores		
Trainable with data	binary or one of N decisions	ranked lists of classes	prob. estimates or continuous scores
No	majority or plurality vote	Borda count	sum, median, or product rules
Yes	weighted vote	logistic regression	Bayes, Dempster-Shafer rules

Because of different contextual requirements, not all methods can be used with all problems. General performance claims about a particular combination strategy are thus difficult to make. Evaluation of the methods is further complicated by the fact that, by and large, only successful experiments are published, and it is difficult to find limits of a method's applicability. Nevertheless, it is obvious that very little can be tuned in the simple voting schemes, and without extensive training data, this may well be about all that can be done. With sophisticated output like estimates of posterior probabilities, more elaborate combination-level classifiers can be applied. However, for problems with a large number of classes, the availability of good estimates of posteriors is a very strong assumption. Without sufficient training data, the estimates given by the individual classifiers are inaccurate, and so are the estimates at the combination level. Thus, applying overly sophisticated combination methods is a dangerous game. The rank combination schemes weaken the requirement to only the availability of preferences, which are always there as long as the classifiers compute any numerical measure of similarity. For very large numbers of classes and mixed type classifiers, this is an interesting middle ground. But the linear scale of the ranks may be too crude an approximation, and to simplify the model, the combinations may have to be restricted to only a small number of top ranks.

The coverage optimization methods suffer less from these problems, as they depend more or less on similar assumptions on the form of component classifiers and decisions. Still, the relative merits of each method are different on each dataset, and it is difficult to predict how a particular method would behave on a problem before actually constructing the classifier collection. As a result, every claim has to rely on empirical testing, and for each new problem the whole trial-and-error process has to be repeated.

Hopes for a better understanding are placed on the development of good theories. But on the theory side, several problems persist. Most theoretical works suffer from a failure to model

various details in a classification problem. For instance, general evidence combination methods, as given in the broader context of artificial intelligence research, are relatively weak because the special structure in the decisions of classifiers are not modeled. There is more knowledge to be exploited than, say, the subjective judgements of human experts.

Numerous arguments on decision combination use some notion of complementariness among the component classifiers. But the precise definition of the complementariness is seldom given. A very common assumption is that the classifiers' decisions are statistically independent, in the sense that the probability of joint success can be modeled as the product of each individual classifier's probability of success. But this is an imposed, very strong assumption which could be very far away from the truth. In a recognition system, classifier decisions are intrinsically correlated, as they respond to the same input. The correlation among the classifiers has to be *measured* from the data. Until measurements confirm the assumption of zero correlation, those theories dependent on it do not necessarily apply. This fact is well known at the level of feature representations, as any one who has compared a decision tree to a simplified Bayes classifier assuming feature independence would know, but is often neglected at the level of classifier decisions.

There is also the dilemma of choosing between a probabilistic view and a geometrical view of classification. Many theories model a classifiers' decision as a probabilistic event, and assume that decisions about each test case are not related to those about others. However, in most application contexts, there is some geometrical continuity in the feature space, such that classifiers' decisions on neighboring points are likely to be similar. Some classifiers, such as decision trees, rely explicitly on this fact to partition the feature space into contiguous regions of the same class. However, with the exception of stochastic discrimination and consistent systems of inequalities, the notion of neighborhood almost never enters decision combination theories. Discussions on the optimal size of component decision trees lightly touch on this, but are never followed through. A precise characterization of the problem geometry will involve descriptors of the fragmentation of the Bayes optimal decision regions, global or local linear separability, and convexity and smoothness of boundaries. Many of these depend on the properties of a specific metric with which the classifier operates. Better modeling of the geometrical behavior of classifiers is attempted in some neural network studies, where classifier training is seen as finding a good approximation to the desired decision boundary, but integration of such models with the probabilistic view is not complete.

The probabilistic view is based on the need to study a problem by random sampling, due to the difficulty of obtaining complete data coverage. Then there are issues of sampling density and sample representativeness, which are intimately related to the classifier's generalization ability [4] [52] and in turn to its accuracy. If the training sample densely covers all the relevant regions in the feature space, many classifiers – as long as they are trainable and expandable in capacity – will work perfectly. Then the competition among methods is more about representation and efficiency. For example, a decision tree may be preferable to nearest neighbors because it is more efficient. So the difficulty of classification is mostly with sparse samples, and for this reason, all theories depending on assumptions of infinite sample size are useless. Those relying on a vague definition of the representativeness of the training samples are not much better, as quite typically such "representativeness" is not even parameterized by the sampling size *relative to the size of the underlying problem* [45].

There are vast differences due to sample size. Consider a space where each point is randomly labeled as one of two classes. Whereas a dense sample may reveal the randomness to some extent, a two-point sample may suggest that the problem is linearly separable. With other less radical problems, sampling density affects the exact ways that isolated subclasses become connected and boundaries constrained, much more than what can be captured in a collective description by a simple count of points. Such problems can occur regardless of the dimensionality of the feature space, though they are more apparent in high-dimensional spaces where the decision boundary can vary with a larger degree of freedom. Observations of empirical evidence [25] [65] [66] suggest strongly that a shortage of samples would ruin most promises from the classical approaches. This fact was addressed in many studies on error rate estimation as well [30] [46] [74].

Vapnik's capacity theory [78] [79] is among the first that directly faces the reality of small sample effects. It provides a link between the interacting effects of classifier geometry and training sample size. But the VC dimension theory is not constructive. It gives only a loose, worst case bound on the error rate given the geometrical characteristics of the classifier and the sample size. The difficulty in tightening the bounds is because of the distribution-free arguments [12] [13]. Nevertheless, as we have seen in the theory of stochastic discrimination, by using a different characterization of training set representativeness, it is possible to show tighter error bounds without involving the VC dimension argument. Also, by using specific geometrical models matched to the problem, it is possible to overcome the infamous *curse of dimensionality* [17].

The theory is difficult because these factors interact with each other. With regard to the problem geometry, the classifier geometry, and the sampling and training processes, what exactly do we mean by saying that two or more classifiers are independent? How about other related concepts such as correlation, diversity, collinearity, coincidence, and equalization? What do they mean in each context where the decisions are represented as one out of many classes, as permutations of class preferences, or as continuous belief scores that are not necessarily probability estimates? Kleinberg's notion of uniformity offered a rigorous definition under the set theoretic abstraction. How can this be generalized to other models of classifier decisions? The bias/variance decomposition [26] gave another way to relate geometry and probabilities, and has been used in analyzing decision boundaries of combined systems [76]. However, such analyses are often flawed due to inadequate assumptions on decision independence.

How can one relate local, point-wise measures of classification error and their correlation to collective measures over the entire training set? How do we relate the error rates of individual classifiers to their degree of agreement? If two classifiers agree only for cases that are easy for both of them, can we still tell whether these two classifiers are independent? How do we know that the agreement is due to the ease of the cases or that the classifiers simply decide by the same mechanism? Detailed studies on the patterns of correlation among the classifiers are necessary to answer these questions [28] [53].

As one compares different approaches of combination, and considers combination of combinations, there are a few more intriguing questions to ask:

- If one defers the final decision and uses the output of individual classifiers only as some scores describing the input, are those scores different in nature from the feature measurements on the input? Are there intrinsic differences between the mappings from the input to the representations given by a classifier versus that of a feature extractor?
- Is combination a necessity or a convenience? Is complementariness intrinsic to some chosen classifier training process? Or is it just an easy way to derive a desired decision boundary?
- Are there any commonalities among all the combination approaches? If many of them are found to be similarly effective, are they essentially doing the same thing despite superficial differences?
- Does the hierarchy of combinations converge to a limit at which one would have exhausted the knowledge contained in the training set, such that no further improvement in accuracy is possible?

6 Precise Characterization of Data Dependences of Performance

Many of the above questions are there because we do not yet have a scientific understanding of the classifier combination mechanisms. We have not taken a close enough look at what is happening in a particular problem. Most of the empirical studies stop on arriving at some collective accuracy claims. But trying a method on 100 published problems and counting how many it wins on does not mean much, because these problems may all be very similar in certain aspects and may not be typical in reality. On the other hand, we will never have a fair sample of realistic problems because this set is hopelessly vague. So what can we do?

If we have a way to characterize the problems that is strongly relevant to classifier accuracy, we may hope to find certain rules relating those characteristics to the behavior of classifiers or systems generated and combined in a specific way. There may be empirical observations that point to opportunities for detailed analysis where theorists can contribute.

Here I am advocating a realist's approach, where the selection of a classifier or a combination method is guided by the characteristics of the data. And the data characteristics must include effects of the problem geometry and sampling density. Statements like "method X is of no help when the training sample is large" are overly simplified. How large is large? An absolute number on the sample size means little without knowledge of the size of a class boundary.

We need much more systematic ways to characterize problems. We need a language to describe problems in ways more relevant to the actions of classifiers, i.e. not merely collective descriptors such as number of classes, number of samples, number of dimensions, etc. We need a better understanding of the geometry and topology of point sets in high dimensional spaces, preservation of such characteristics under feature transformations and sampling processes, and their interaction with the primitive geometrical models used in most well known classifiers. We need to measure or estimate the length and curvature of the class boundaries, fragmentation of the decision regions in terms of existence, size, and connectedness of subclasses, and the stability of these characteristics as the sampling density changes. Some recent attempts are possible starting points [16] [38] [39] [40] [70].

Once we find ways to characterize the problems, we can then ask the question: what type of problems does this method work for? We have to accept that a probabilistic flavor will remain whenever we have to deal with unseen data. This uncertainty is intrinsic. It can only be reduced by knowledge of the structural regularity of the problem, but can never be removed.

Given a problem in a fixed feature space, is there a limit on what the automatically trainable methods can do? Recall that all such methods are based on certain particular geometrical primitives, such as convex regions, axis parallel cuts, rectangular boxes, Gaussian kernels, piecewise linear boundaries, etc. It needs to be established that such models will fit decision regions of arbitrary shapes and degrees of connectedness. At which point should we say that it is meaningless to continue training, and that any more improvement in accuracy will be from luck rather than effort? VC dimension theories give us a certain limit, but it could be very remote from what is achievable. By exploiting the structural regularities of the problems and matching them to appropriate classifiers, we should be able to do better than that.

7 Conclusions

I reviewed several lines of ideas developed in the past decade on methods for combining parallel classifiers. In summary, the main results include

1. Numerous empirical, quantitative results that accuracies of combined systems can be better than that of each individual.
2. Dozens of practical algorithms for combining decisions by a small number of classifiers.
3. Several ways to generate a collection of classifiers containing complementary strengths.
4. A formal analysis of the relationship between individual classifier accuracy and generalization ability, their mutual agreement, and combined accuracy.

Among the many open questions raised in the studies, those involving the interaction of the geometric and probabilistic characteristics of a problem are the most intriguing. They hold the key to a better understanding and further improvements of the methods. An essential need in this direction is to find a better set of descriptors for the structure of a problem in the feature space and to describe the behavior of classifiers in corresponding terms. These descriptors can be used to categorize real world problems, which would permit studies and prediction of the behavior of various classifiers and combination methods on a whole class of problems.

Besides detailed studies of the behavior of each combination technique and their interactions with problem characteristics, several methodological directions are also worth pursuing:

- Ingenious designs of feature extractors and similarity measures that can simplify the class boundary will continue to play an important role in real applications. Systematic searches with the same goal are even more interesting.
- Unsupervised learning will play an increasing role in the context of supervised learning. Clustering methods will be applied more extensively and systematically to better understand the geometry of the class boundaries and their sensitivity to sampling density. Other ways for describing data, such as estimation of intrinsic and extrinsic dimensionalities of the classes or subclasses, or probabilistic mixture decomposition models, will also be helpful.
- More emphasis should be put on localized (or dynamically selected) classification methods. A blind application of everything to everything will prove to be inferior to localized methods. Systematic strategies should be developed to fine tune the classifiers to the characteristics of local regions and to associate them with corresponding input.
- A better understanding is needed about the differences between deterministic and stochastic classifier generating methods. This will need a careful study of the exact role of randomization in various classifier or combination tuning processes, and the corresponding geometrical effects.
- New methods can come from a merger of the decision optimization and coverage optimization strategies, such that collections of fixed classifiers can be enhanced by introducing additional components with enforced complementariness, and coverage optimization methods may use more sophisticated decision combination schemes.

By now, classifier combination has become a rich and exciting area with much proven success. This review is biased towards criticisms rather than celebrations of the ideas explored over its brief history. It is my hope that these discussions can call attention to some of the confusion and missing links in the methodology, and point out the more fruitful directions for further research. We can see that some of the recent developments are already moving towards these directions. This is very encouraging.

Acknowledgements

The author wishes to thank George Nagy and the anonymous reviewers for detailed comments on the manuscript, and Eugene Kleinberg for many discussions over the past ten years on the theory of stochastic discrimination, its comparison to other approaches, and perspectives on the fundamental issues in pattern recognition.

References

[1] K.J. Arrow, *Social Choice and Individual Values*, John Wiley & Sons, Inc., New York, 1951, 2nd ed., 1963.

[2] J.A. Benediktsson, "Consensus theoretic classification methods," *IEEE Transactions on Systems, Man, and Cybernetics*, **SMC-22**, 4, July/August 1992, 688-704.

[3] C. Berenstein, L.N. Kanal, D. Lavine, "Consensus and evidence," in E.S. Gelsema, L.N. Kanal, (eds.), *Pattern Recognition and Artificial Intelligence II*, North Holland, 1986, 523-546.

[4] R. Berlind, *An Alternative Method of Stochastic Discrimination with Applications to Pattern Recognition*, Doctoral Dissertation, Department of Mathematics, State University of New York at Buffalo, 1994.

[5] D. Black, *The Theory of Committees and Elections*, Cambridge University Press, London, 1958, 2nd ed., 1963.

[6] K.P. Bogart, "Preference Structures I: Distances between Transitive Preference Relations," *Journal of Mathematical Sociology*, **3**, 1973, 49-67.

[7] K.P. Bogart, "Preference Structures II: Distances Between Asymmetric Relations", *SIAM Journal of Applied Mathematics*, **29**, 2, September 1975, 254-262.

[8] L. Breiman, "Bagging predictors," *Machine Learning*, **24**, 1996, 123-140.

[9] H. Bunke, "Hybrid methods in pattern recognition," in P.A. Devijver, J. Kittler, (eds.), *Pattern Recognition Theory and Applications*, Springer-Verlag, 1987, 367-382.

[10] N.C. de Condorcet, *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix.* Imprimerie Royale, Paris, 1785.

[11] B.V. Dasarathy, B.V. Sheela, "A composite classifier system design: concepts and methodology," *Proceedings of the IEEE*, **67**, 5, May 1979, 708-713.

[12] L. Devroye, "Any discrimination rule can have an arbitrarily bad probability of error for finite sample size," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **4**, 2, March 1982, 154-157.

[13] L. Devroye, "Automatic pattern recognition: a study of the probability of error," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **10**, 4, July 1988, 530-599.

[14] T.G. Dietterich, "Machine-learning research: four current directions," *AI Magazine*, **18**, 4, Winter 1997, 97-135.

[15] T.G. Dietterich, G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," *Journal of Artificial Intelligence Research*, **2**, 1995, 263-286.

[16] R.P.W. Duin, "Compactness and Complexity of Pattern Recognition Problems," in C. Perneel, (eds.), *Proc. Int. Symposium on Pattern Recognition, In Memoriam Pierre Devijver*, Royal Military Academy, Brussels, Feb 12, 1999, 124-128.

[17] R.P.W. Duin, "Classifiers in almost empty spaces," *Proceedings of the 15th International Conference on Pattern Recognition*, Barcelona, September 3-8, 2000, **II**, 1-7.

[18] J.L. Engvall, "A least upper bound for the average classification accuracy of multiple observers," *Pattern Recognition*, **12**, 415-419.

[19] P.C. Fishburn, *The Theory of Social Choice*, Princeton University Press, Princeton, 1972.

[20] P.C. Fishburn, *Interprofile Conditions and Impossibility*, Fundamentals of Pure and Applied Economics 18, Hardwood Academic Publishers, 1987.

[21] Y. Freund, R.E. Schapire, "Experiments with a New Boosting Algorithm," *Proceedings of the Thirteenth International Conference on Machine Learning*, Bari, Italy, July 3-6, 1996, 148-156.

[22] J. Friedman, T. Hastie, R. Tibshirani, "Additive logistic regression: a statistical view of boosting," *Annals of Statistics*, **28**, 2, April 2000, 337-374.

[23] K.S. Fu, *Syntactic Pattern Recognition and Applications*, Prentice-Hall, 1982.

[24] K.S. Fu, "A step towards unification of syntactic and statistical pattern recognition," *IEEE Transactions of Pattern Analysis and Machine Intelligence*, **PAMI-5**, March 1983, 200-205.

[25] K. Fukunaga, D.L. Kessell, "Estimation of classification error," *IEEE Transactions on Computers*, **20**, 12, December 1971, 1521-1527.

[26] S. Geman, E. Bienenstock, R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Computation*, **4**, 1992, 1-58.

[27] D. Gernert, "Distance or similarity measures which respect the internal structure of objects," *Methods of Operations Research*, **43**, 1981, 329-335.

[28] G. Giacinto, F. Roli, "A theoretical framework for dynamic classifier selection," *Proceedings of the 15th International Conference on Pattern Recognition*, Barcelona, September 3-8, 2000, **II**, 8-11.

[29] L.A. Goodman, H. Markowitz, "Social welfare functions based on individual rankings," *The American Journal of Sociology*, **58**, 1952, 257-262.

[30] D.J. Hand, "Recent advances in error rate estimation," *Pattern Recognition Letters*, **4**, October 1986, 335-346.

[31] L.K. Hansen, P. Salamon, "Neural network ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-12**, 10, October 1990, 993-1001.

[32] L.V. Hedges, I. Olkin, *Statistical Methods for Meta-Analysis*, Academic Press, 1985.

[33] T.P. Hettmansperger, *Statistical Inference Based on Ranks*, John Wiley & Sons, 1984.

[34] T.K. Ho, J. J. Hull, S.N. Srihari, "Decision Combination in Multiple Classifier Systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-16**, 1, January 1994, 66-75.

[35] T.K. Ho, "Adaptive coordination of multiple classifiers," in J.J. Hull, S.L. Taylor (eds.), *Document Analysis Systems II*, World Scientific Publishing Co., 1997, 371-384.

[36] T.K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**, 8, August 1998, 832-844.

[37] T.K. Ho, "Nearest neighbors in random subspaces," *Proceedings of the Second International Workshop on Statistical Techniques in Pattern Recognition*, Sydney, Australia, August 11-13, 1998, 640-648.

[38] T.K. Ho, M. Basu, "Measuring the Complexity of Classification Problems," *Proceedings of the 15th International Conference on Pattern Recognition*, Barcelona, September 3-8, 2000, **II**, 43-47.

[39] T.K. Ho, "Complexity of classification problems and comparative advantages of combined classifiers," in J. Kittler, F. Roli, (eds.), *Multiple Classifier Systems*, Lecture Notes in Computer Science 1857, Springer, 2000, 97-106.

[40] A. Hoekstra, R.P.W. Duin, "On the nonlinearity of pattern classifiers," *Proc. of the 13th ICPR*, Vienna, August 1996, **D**271-275.

[41] Y.S. Huang, C.Y. Suen, "A method of combining multiple experts for the recognition of unconstrained handwritten numerals," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-17**, 1, January 1995, 90-94.

[42] A.K. Jain, S. Prabhakar, S. Chen, "Combining multiple matchers for a high security fingerprint verification system," *Pattern Recognition Letters*, **20**, 1999, 1371-1379.

[43] A.K. Jain, R.P.W. Duin, J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-22**, 1, January 2000, 4-37.

[44] C. Ji, S. Ma, "Combinations of weak classifiers," *IEEE Transactions on Neural Networks*, **8**, 1, January 1997, 32-42.

[45] L. Kanal, B. Chandrasekaran, "On dimensionality and sample size in statistical pattern classification," *Pattern Recognition*, **3**, 1971, 225-234.

[46] J. Kittler, P.A. Devijver, "Statistical properties of error estimators in performance assessment of recognition systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **4**, 2, March 1982, 215-220.

[47] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-20**, 3, March 1998, 226-239.

[48] J. Kittler, F. Roli, (eds.), *Multiple Classifier Systems*, Lecture Notes in Computer Science 1857, Springer, 2000.

[49] E.M. Kleinberg, "Stochastic Discrimination," *Annals of Mathematics and Artificial Intelligence*, **1**, 1990, 207-239.

[50] E.M. Kleinberg, "An overtraining-resistant stochastic modeling method for pattern recognition," *Annals of Statistics*, **4**, 6, December 1996, 2319-2349.

[51] E.M. Kleinberg, "On the algorithmic implementation of stochastic discrimination," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-22**, 5, May 2000, 473-490.

[52] E.M. Kleinberg, "A mathematically rigorous foundation for supervised learning," in J. Kittler, F. Roli, (eds.), *Multiple Classifier Systems*, Lecture Notes in Computer Science 1857, Springer, 2000, 67-76.

[53] L.I. Kuncheva, C.J. Whitaker, C.A. Shipp, R.P.W. Duin, "Is independence good for combining classifiers?" *Proceedings of the 15th International Conference on Pattern Recognition*, Barcelona, September 3-8, 2000, **II**, 168-171.

[54] L. Lam, "Classifier combinations: implementations and theoretical issues", in J. Kittler, F. Roli, (eds.), *Multiple Classifier Systems*, Lecture Notes in Computer Science 1857, Springer, 2000, 77-86.

[55] L. Lam, C.Y. Suen, "Application of majority voting to pattern recognition," *IEEE Transactions on Systems, Man, and Cybernetics*, **SMC-27**, 5, September/October 1997, 553-568.

[56] M. LeBlanc, R. Tibshirani, "Combining estimates in regression and classification," *Journal of the American Statistical Association*, **91**, 436, December 1996, 1641-1650.

[57] D.S. Lee, *A Theory of Classifier Combination: The Neural Network Approach*, Doctoral Dissertation, Department of Computer Science, State University of New York at Buffalo, 1995.

[58] R.C. Luo, "Multisensor integration and fusion in intelligent systems," *IEEE Transactions on Systems, Man, and Cybernetics*, **SMC-19**, 5, September/October 1989, 901-931.

[59] E. Mandler, J. Schuermann, "Combining the classification results of independent classifiers based on Dempster/Shafer theory of evidence," in E.S. Gelsema, L.N. Kanal, (eds.), *Pattern Recognition and Artificial Intelligence*, North Holland, 1988, 381-393.

[60] V.D. Mazurov, A.I. Krivonogov, V.L. Kazantsev, "Solving of Optimization and Identification Problems by the Committee Methods," *Pattern Recognition*, **20**, 4, 1987, 371-378.

[61] R. Meddis, *Statistics Using Ranks, A Unified Approach*, Basil Blackwell, 1984.

[62] S.J. Nowlan, "Competing experts: an experimental investigation of associative mixture models," Technical Report CRG-TR-90-5, Department of Computer Science, University of Toronto, September 1990.

[63] D. Partridge, W.B. Yates, "Engineering multiversion neural-net systems," *Neural Computation*, **8**, 4, 1996, 869-893.

[64] L.F. Pau, "Fusion of multisensor data in pattern recognition," in J. Kittler, K.S. Fu, and L.F. Pau, (eds.), *Pattern Recognition Theory and Applications*, Reidel, 1982, 189-201.

[65] S. Raudys, V. Pikelis, "On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **2**, 3, May 1980, 242-252.

[66] S. Raudys, A.K. Jain, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **13**, 3, 1991, 252-264.

[67] R.E. Schapire, Y. Freund, P. Bartlett, W.S. Lee, "Boosting the margin: a new explanation for the effectiveness of voting methods," *Annals of Statistics*, **26**, 5, October 1998, 1651-1686.

[68] G. Shafer, *A Mathematical Theory of Evidence*, Princeton University Press, 1976.

[69] A.J.C. Sharkey, *Combining Artificial Neural Nets: Ensemble and Modular Multi-Net Systems*, Springer-Verlag, 1999.

[70] S.Y. Sohn, "Meta analysis of classification algorithms for pattern recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **21**, 11, 1999, 1137-1144.

[71] M. Skurichina, R.P.W. Duin, "Boosting in linear discriminant analysis," in J. Kittler, F. Roli, (eds.), *Multiple Classifier Systems*, Lecture Notes in Computer Science 1857, Springer, 2000, 190-199.

[72] S.N. Srihari, "Reliability analysis of majority vote systems," *Information Sciences*, **26**, 1982, 243-256.

[73] S.N. Srihari, "Reliability analysis of biased majority-vote systems," *IEEE Transactions on Reliability*, **R-31**, 1, April 1982, 117-118.

[74] G.T. Toussaint, "Bibliography on estimation of misclassification," *IEEE Transactions on Information Theory*, **20**, 4, July 1974, 472-479.

[75] J.D. Tubbs, W.O. Alltop, "Measures of confidence associated with combining classification results," *IEEE Transactions on Systems, Man, and Cybernetics*, **SMC-21**, 3, May/June 1991, 690-692.

[76] K. Tumer, J. Ghosh, "Analysis of decision boundaries in linearly combined neural classifiers," *Pattern Recognition*, **29**, 1996, 341-348.

[77] K. Tumer, J. Ghosh, "Linear and order statistics combiners for pattern recognition," in A. Sharkey, (ed.), *Combining Artificial Neural Nets: Ensemble and Modular Multi-Net Systems*, Springer-Verlag, 1999, 127-162.

[78] V. Vapnik, *Estimation of Dependences Based on Empirical Data*, Springer-Verlag, 1982.

[79] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, 1998.

[80] K.D. Wernecke, "A coupling procedure for the discrimination of mixed data," *Biometrics*, **48**, June 1992, 497-506.

[81] D.H. Wolpert, "Stacked generalization," *Neural Networks*, **5**, 1992, 241-259.

[82] K. Woods, W.P. Kegelmeyer Jr., K. Bowyer, "Combination of multiple classifiers using local accuracy estimates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-19**, 4, April 1997, 405-410.

[83] L. Xu, A. Krzyzak, C.Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Transactions on Systems, Man, and Cybernetics*, **SMC-22**, 3, May/June 1992, 418-435.