Data Complexity Analysis for Classifier Combination

Tin Kam Ho

Bell Laboratories, Lucent Technologies
700 Mountain Ave., 2C-425, Murray Hill, NJ 07974, USA

Abstract. Multiple classifier methods are effective solutions to difficult pattern recognition problems. However, empirical successes and failures have not been completely explained. Amid the excitement and confusion, uncertainty persists in the optimality of method choices for specific problems due to strong data dependences of classifier performance. In response to this, I propose that further exploration of the methodology be guided by detailed descriptions of geometrical characteristics of data and classifier models.

1 Introduction

Multiple classifier systems are often practical and effective solutions for difficult pattern recognition tasks. The idea appeared in 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. In other areas it occurred naturally from the application context, such as the need to employ a variety of sensor types which induces a natural decomposition of the problem. There were also cases motivated by an attempt to escape from the burden of making a commitment to some arbitrary initial condition, such as the initial weights for a neural network. There were even hopes that any sort of randomness introduced in classifier training would produce a diverse collection that could 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 on which each specific classifier performs well, and direct new input accordingly. A sequential approach would use one classifier first, and invoke others only if it fails to yield a decision with sufficient confidence. All these can be said to be multiple classifier strategies, and have been explored to a certain extent. However, motivated by the above mentioned factors, most 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 alternative approaches?

The trend of parallel combination of many classifiers deviates from, or even follows an opposite philosophy of, the traditional selection approach in which one

evaluates the available classifiers against a representative sample and chooses the best one to use. Here, in essence, one abandons the attempt to find the best classifier, and instead, tries to use all the available ones in a smart way. This opposes the wisdom of economical design. But introducing needless classifiers is more than just harming efficiency. The agreement of two poor classifiers do not necessarily yield more correct decisions. And one can easily 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.

As the idea prospered, many different proposals of combination 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 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 graduately losing insight into the original problem.

So, is classifier combination a well justified, systematic methodology, or is it a desperate attempt to make the most out of imperfect designs? What exactly is gained or lost in a combination effort? What has been achieved and what is still missing, and what should we do next? In this paper, I review the proposed methods and some of the challenges in related theories and practices, and suggest ways to further advance the methodology.

2 Difficulties in Combination Theories and Practices

The possibility, by now well supported by empirical evidences, of being able to go beyond the power of traditional classifiers is exciting. In pattern recognition, early discoveries that the combined accuracy of several classifiers can be better than that of each individual came as a surprise from experiments. Later studies revealed many alternative methods that can achieve similar effects. Proposed methods fall into two categories: (1) assume a given, fixed set of carefully designed and highly specialized classifiers, attempt to find an *optimal combination* of their decisions; and (2) assume a fixed decision combination function, *generate* a set of mutually complementary, generic classifiers that can be combined to achieve optimal accuracy. We will refer to combination strategies of the first kind as *decision optimization* methods and the second kind as *coverage optimization* methods. It is also possible to apply the decision optimization methods to classifiers generated with the aim of coverage optimization.

Decision Optimization

Often in pattern recognition practices, several classifiers can be designed for the same problem. Combining their decisions gives opportunities of improving

accuracy or reliability. 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 suitable under two contextual requirements.

Table 1. Decision combination methods

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

The idea of employing multiple experts specialized for a given task in different aspects is probably as old as the history of human society. But such common wisdom does not necessarily and immediately apply in the context of pattern recognition, where concepts such as differences and cooperation have specific meanings. Behavior of classifiers can be mathematically characterized, and with accuracy being an objective measure of effectiveness, the benefits of any combination method can potentially be quantified. Reasons of 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 early explorations show that there are vast differences in the effectiveness of different combination procedures.

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 in lack of sufficient training data, this may well be about all that can be done. With sophisticated output like estimates of posterior probabilities, other than the simple sum, product rules or Bayes schemes, more elaborated 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 combining schemes weaken the requirement to only preferences which are always available

as long as the classifiers compute any numerical similarity measure. For a very large number of classes and mixed type classifiers, this is an interesting middle ground. But the linear and uniform 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.

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 failures to model various details in a classification problem.

In behavior sciences, methods for combining votings and rankings of alternative choices are referred to as *social choice functions* or *group consensus functions*. There, 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, prior performance of the voters can also be evaluated against the objective truth, and based on such evaluation the combination function can be tuned. Some combination schemes such as regression make explicit use of these evaluations. Others that do not use such information carry an implicit assumption that the voters are competent to a certain extent and that the imposed characterization of the voters' preference is reasonable, either of which may not be correct.

Moreover, 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. These two effects must be modeled separately.

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 dominated. Prior performances of individual classifiers can be embedded into the combination function in the form of estimates of correctness probabilities conditioned on the individual decisions. Simpler combination rules that do not take into account the classifiers' prior performance were studied in [17], where a justification was given in support of 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 explored 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 classification principles. In statistics they are known as model-mix methods [20], justified by reducing the bias of the combined estimator.

Many attempts in modeling the performance of such combination schemes use some notion of complementariness among the component classifiers. But the

precise definition of it is seldom given. A very common assumption is that the classifiers' decisions are statistically independent, in the sense that the probability of a joint decision equals the product of probabilities of each individual classifier's corresponding decision. 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 aware of at the level of feature representations, but is often neglected at the level of classifier decisions. Moreover, the correlation among different classifiers' decisions can vary from input to input. Decisions may be strongly corrected only on easy cases far from the class boundary. So even with the same set of classifiers, correlation of their decisions varies across problems and subproblems according to proportions of easy and hard cases. It is a theoretical challenge to model such detailed data dependency.

Coverage Optimization

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' decision is sufficiently close to being correct. Coverage optimization methods are called on to pursue the missing guarantee. There, 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. Table 2 summarizes the better known coverage optimization methods and the training mechanism used to introduce complementariness between the components.

Table 2. Coverage optimization methods

Method	Training mechanism for introducing complementariness
perturbation	vary initial conditions or parameters of training process
stacking	train classifiers by nonoverlapping subsamples of training set
bagging	resample the training set by bootstrap replicates
boosting	resample the training set by weights evolving with accuracy
random subspaces	project training set to randomly chosen subspaces
stochastic discrimination	generate random kernels to measure coverage of training set
error correction output coding	force training on partial decision boundaries

Several methods introduce complementary strengths by training component classifiers on different subsamples of the training set. Despite many observed empirical successes, such training set subsampling methods are paradoxical. Weakening the individual classifiers by not training or equally weighing on all avail-

able data is said to help avoid overfitting. At an extreme, boosting 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 it or not that the classifiers perfectly adapt to the training data? If we want it, what is the point of deferring it to the level of decision combination? If we do not want it, 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 way followed by the component classifiers? If the training set is assumed to represent well the unseen cases, why would one believe that by sacrificing training set accuracy one can gain in 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, there is no reason to bother with such elaborated 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 evidences do 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, details matter on the specific way data are split at each internal node. 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 to some success. With others, like linear discriminators, things are less clear. If the component classifiers are too weak, given the simple decision combination function involving weighted or unweighted averaging, the decisions of many bad classifiers can easily outweigh the good ones, especially in methods like boosting that focus more and more on the errors. And how about mixing in different types of classifiers?

Such fundamental issues are in the midst of confusion in several communities. In [18] Kleinberg offered a rigorous analysis of these issues, using a set-theoretic abstraction to remove all the algorithmic details of classifiers, feature extractors, 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 into the power set of the feature space. If the sample satisfies a uniformity condition, i.e., if its coverage is unbiased to 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. Then it is shown that discrimination between classes is possible as long as there is 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 weak models, i.e., if each model is thick enough with respect to the spatial continuity of the classes, 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 approximation of the estimate, which will translate, by the symmetry, to accuracy in classifying unseen points.

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 projectability of the weak models is a key element in the proof, and not an implicit side effect, as in several other theories attempted to explain the behavior of the coverage optimization methods. 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 subjects them less to sampling errors carried in small training sets [27] [28]. 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 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 succeeded in offering a complete explanation of the combined behavior of such simple classifiers. However, much remains to be explored in using it to predict the behavior of fewer but more sophisticated classifiers.

Combination theories must deal with the dilemma of choosing between a probabilistic view and a geometrical view of classification, or the difficulty of blending the two. Many theories model a classifiers' decision as a probabilistic event, and assume that decisions on each input are not related to decisions on 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. But the notion of neighborhood is explictly used only in a few theories such as stochastic discrimination and consistent systems of inequalities [21]. Discussions on the optimal size of component decision trees barely touch on this, but are not followed through. Precise characterization of the problem geometry will involve descriptions of the fragmentation of the Bayes optimal decision regions, global or local linear separability, and convexity and smoothness of boundaries. Many of these depend the properties of a specific metric based on 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 comes in because of the need to study a problem by random sampling due to the unavailability of complete data coverage. Then there is the issue of sampling density and training sample representativeness, which is intimately related to the classifier's generalization ability, 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, and then the competition among methods is more on representation and execution efficiency. Say, a decision tree may be preferable to nearest neighbors for efficiency reasons. So the difficulty of classification is mostly with sparse samples – and for this reason, all theories depending on assumptions of infinite samples are useless. Those relying on a vague definition of representativeness of the training samples is not much better, as quite typically such representativeness is not even parameterized by the sampling size *relative to the size of the underlying problem.*

There are vast differences due to sampling density. 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 are constrained, much more than what can be captured in a collective description by a single 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 evidences [7] [22] [23] suggested strongly that shortage of samples would ruin most promises from the classical approaches. This fact was addressed in many studies on error rate estimation as well [11] [16] [25].

Vapnik's capacity theory [27] [28] 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 [1] [2]. Nevertheless, as we have seen in the theory of stochastic discrimination, with 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* [4].

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 offers a rigorous definition under the set-theoretic abstraction. How can this be generalized to other models of classifier decisions? The bias/variance decomposition [8] gives another way to relate geometry and probabilities, and has been used in analyzing decision boundaries of combined systems [26]. Though, many analyses are problematic 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? Say, if we observe two classifiers agreeing 99% of the time and differing for the rest, how much can we infer about the overall similarity of their decisions? And how likely is the agreement observed in a different problem? If all the agreed cases are correct decisions, simply because those cases are easy for both classifiers, can we tell anything about whether the classifiers decide by the same mechanism? Detailed studies on the patterns of correlation among the classifiers are necessary to answer these questions [10] [19].

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 just 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 or by a feature extractor?
- Is combination a necessity or a convenience? Is there some complementariness intrinsic to certain classifier training processes? 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 upon which one would have exhausted the knowledge contained in the training set, such that no further improvement in accuracy is possible?

3 Precise Characterization of Data Dependences of Performance

Many of the above questions are there because we do not yet have a detailed, scientific understanding of the classifier combination mechanisms. Theoretical explanations are often incomplete, or they have to stop at a level where the combinatorics defy detailed modeling. Many studies attempt to analyze classifier behavior for all possible problems and data distributions, which result inevitably in very weak performance bounds. On the other hand, empirical evidences are

often very specific to particular applications, and are seldom systematically documented. Attempted comparative studies often stop on arriving at some collective accuracy measures. But trying a method on 100 published problems and counting how many it wins on does not mean much, because these problem 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 of the ill-definition of the set. So what can we do?

If we have a way to characterize the problems in good relevance to classifier accuracy, we may hope to find certain rules relating those characteristics to the behavior of classifier or systems generated and combined in a specific way. Empirical observations of such relationships may point to opportunities for detailed analysis of underlying reasons.

Here I am advocating a realist's approach where selection of a classifier or a combination method is guided by 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 simplifying. How large is large? An absolute number on the sample size means little without knowledge of the length of class boundary.

We need much more systematic ways to characterize the problems. We need a language to describe the problems in aspects 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 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 with respect to changes in sampling density. Some recent attempts are interesting starting points [3] [12] [13] [14] [24].

Characterization of Data Complexity

In reality, most practical classification problems arise from nonchaotic processes many of which can be described by an underlying physical model. Though the models may contain a stochastic component, there should still exist a significant structure in the resulting class distributions that differs from a random labeling of points. An analysis of such differences will provide us with a framework in which one can study the behavior of specific classifiers and combination methods.

Structured data differ from random labeling in that with random labeling there is no geometrical continuity or regularity based on which inferences can be made about labels of unseen points from the same source. On the other hand, automatic classification methodologies are based on the assumption that such learning is possible, to various degrees of difficulty.

Obviously one practical measure of problem difficulty is the error rate of a chosen classifier. However, since our eventual goal is to study behavior of clas-

sifiers, other measures should also be explored that are independent of such choices. Moreover, a problem can be difficult for different reasons. Certain problems are known to have nonzero Bayes error. There the classes are ambiguous either intrinsically or due to inadequate feature measurements. Others may have a complex class boundary and/or subclass structures. Sometimes high dimensionality of the feature space and sparseness of available samples are to be blamed.

Among the different reasons, the geometrical complexity of the class boundaries is probably most ready for detailed investigation. One can choose the class boundary to be the simplest (of minimum measure in the feature space) decision boundary that minimizes Bayes error. With a complete sample, the class boundary can be characterized by its Kolmogorov complexity. A problem is complex if it takes a long algorithm (possibly including an enumeration of all the points and their labels) to describe the class boundary. This aspect of difficulty is due to the nature of the problem and is unrelated to the sampling process. Kolmogorov complexity describes the absolute amount of information in a dataset, and is not algorithmically computable. By geometrical complexity one focuses on descriptions of regularities and irregularities contained in the dataset in terms of geometrical primitives. This would be sufficient for pattern recognition where classifiers can also be characterized by similar geometrical terms.

An incomplete or sparse sample adds another layer of difficulty to a discrimination problem, since an unseen point in the vicinity of some training points may share their class labels according to different generalization rules. In real world situations, often a problem becomes difficult because of a mixture of these two effects. Sampling density is more critical for an intrinsically complex problem than an intrinsically simple problem (e.g. a linearly separable problem with wide margins). If the sample is too sparse, an intrinsically complex problem may appear deceptively simple. Thus, in lack of a complete sample, such measures of the problem complexity have to be qualified by the representativeness of the training set.

Several measures of geometrical complexity were studied in [12], where it is shown that many real-world problems occupy a continuum between two extremes given by random labelings (the most difficult problems) and linearly separable problems (the easiest). Measures known to be useful fall into several groups, characterizing the linearity of decision boundaries, inter- and intra- class point proximity, overlap of class-specific convex hulls and their projections into subspaces, and existence and connectivity of subclasses. In addition, interesting characteristics of the datasets were observed from the behavior of two primitive classifiers, i.e, nearest neighbor and linear discriminant (obtained by linear programming minimizing sum of distances of error points to the separating hyperplane). Their error rates and measures of the intersection of their error points with the class-specific convex hulls give some hints to the problem geometry. Table 3 summarizes the definitions of these measures and the effects described by each. Details can be found in [12] and [13].

Some of these measures are correlated and can be indications of more fundamental geometrical or topological characteristics. Furthermore, sensitivity of

Table 3. Measures of geometrical complexity

Relevant effects	Measures
clustering and sphericity	% points with associated adherence subsets retained
inter versus intra class proximity	% points on boundary expressed as between-class edges in a minimum spanning tree
	error rate of nearest neighbor classifier
overlap of classes	maximum Fisher's discriminant ratio
	maximum (individual) feature efficiency
	ratio of average intra/inter class nearest neighbor distance
	volume of overlap region
linear separability	LP minimized sum of error distances
	error rate of linear classifier by LP
curvature of boundary	nonlinearity of linear classifier by LP
	nonlinearity of nearest neighbor classifier

these measures with respect to the sampling density is also important. This can potentially be estimated by repeating the measurements on subsamples of the given data set.

Coupling between Data Characteristics and Classifier Models

Once we find ways to characterize the problems, we can then ask the question: what type of problems does a particular classifier or combination method work for?

To investigate the optimality of match between a classifier or combination method and a given problem, one needs a detailed characterization of the geometrical structure of the decision regions given by the classifier, and the modifications introduced by the combination method. For example, decision trees splitting on single features divide the feature space into a set of hyper-rectangular cells within each classification remains invariant. Voting of two such trees further segments those cells into the cross-product of the two sets. How many of the boundary faces of neighboring cells are close to the class boundary depends on the geometry of the problem – whether the problem has axis-parallel, flat, or curved boundary surfaces, and disconnected subclasses, and to what extent.

An example is given in [13] where the performances of two decision forest constructors, namely, bagging and random subspaces, were studied along with several measures of problem geometry. Strong correlation was observed between the classifier accuracies and a measure of length of class boundaries as well as a measure of the thickness of class manifolds. Follow-up studies showed that both types of forests are capable of improving over a single tree for problems of various levels of complexity, i.e., improvements are observed in problems with very low, very high, and all intermediate single tree error rates. Neither works well when several conditions occur jointly: a very high fraction of boundary points (70% or above), a close to 1 ratio of intra/inter class nearest neighbor distances, a very low maximum Fisher's discriminant ratio (0.05 or below), and high nonlinearity

of both nearest neighbor and linear classifiers (25% or above). For these cases a single tree performs poorly, nor will forest methods offer much help. Easier cases are those with relatively compact classes, i.e., when the pretopological measure is lower than 80%. For such cases improvements by forests over single trees are less significant. The geometrical measures also reveal comparative advantages of each type of forests. The subsampling method is preferable when the training set is very sparse relative to dimensionality, especially when coupled with a close-to-vanishing maximum Fisher's ratio (0.3 or below), and when the class boundary is highly nonlinear. Subspace forests perform better when the class boundary is smoother (both nearest neighbor and linear classifiers display low nonlinearity). If the training set is large relative to dimensionality, the subspace method is more preferable, even if the class distributions are long and thin. Rules similar to these may be observed in further studies.

Given a problem in a fixed feature space, is there a limit on how well an automatically trainable classifier 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 into arbitrary shaped decision regions with arbitrary degree 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 for certain classes of problems, by exploiting the structural regularities and matching them to appropriate classifiers, we should be able to do better than that. Knowledge on the structural regularity may not be sufficient to remove entirely the probabilistic nature of the estimates inherent in unseen data, but should help in reducing the level of uncertainty.

4 Conclusions

I reviewed some challenges in the theories and practices of combining parallel classifiers. Many open questions point to a lack of insight into the intriguing interactions of the geometric and probabilistic characteristics of a problem and the classifier models. A thorough understanding of such interactions holds the key to further improvements of the methods. An essential need in this direction is to find a better set of descriptors for the geometrical 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 the real world problems and their subproblems or transformations, which would permit studies and prediction of the behavior of various classifiers and combination methods on a whole class of cases.

In addition, 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 its sensitivity to sampling density. Others such as estimation of intrinsic and extrinsic dimensionalities of the classes or subclasses 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.
- Better understanding is needed to choose 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 merging the decision optimization and coverage optimization strategies, such that collections of specialized 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. It is my hope that these discussions can call for attention to some of the confusions and missing links in the methodology, and point out the more fruitful directions for further research. Some recent developments are already moving towards these directions. This is very encouraging.

References

1. 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.
2. 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.
3. 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.
4. 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.
5. J.L. Engvall, A least upper bound for the average classification accuracy of multiple observers, *Pattern Recognition*, **12**, 415-419.
6. P.C. Fishburn, *The Theory of Social Choice*, Princeton University Press, Princeton, 1972.
7. K. Fukunaga, D.L. Kessell, Estimation of classification error, *IEEE Transactions on Computers*, **20**, 12, December 1971, 1521-1527.
8. S. Geman, E. Bienenstock, R. Doursat, Neural networks and the bias/variance dilemma, *Neural Computation*, **4**, 1992, 1-58.

9. D. Gernert, Distance or similarity measures which respect the internal structure of objects, *Methods of Operations Research*, **43**, 1981, 329-335.
10. 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.
11. D.J. Hand, Recent advances in error rate estimation, *Pattern Recognition Letters*, **4**, October 1986, 335-346.
12. 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.
13. 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.
14. A. Hoekstra, R.P.W. Duin, On the nonlinearity of pattern classifiers, *Proc. of the 13th ICPR*, Vienna, August 1996, **D**271-275.
15. 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.
16. 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.
17. 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.
18. E.M. Kleinberg, An overtraining-resistant stochastic modeling method for pattern recognition, *Annals of Statistics*, **4**, 6, December 1996, 2319-2349.
19. 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.
20. M. LeBlanc, R. Tibshirani, Combining estimates in regression and classification, *Journal of the American Statistical Association*, **91**, 436, December 1996, 1641-1650.
21. 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.
22. 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.
23. 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.
24. S.Y. Sohn, Meta analysis of classification algorithms for pattern recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **21**, 11, 1999, 1137-1144.
25. G.T. Toussaint, Bibliography on estimation of misclassification, *IEEE Transactions on Information Theory*, **20**, 4, July 1974, 472-479.
26. K. Tumer, J. Ghosh, Analysis of decision boundaries in linearly combined neural classifiers, *Pattern Recognition*, **29**, 1996, 341-348.
27. V. Vapnik, *Estimation of Dependences Based on Empirical Data*, Springer-Verlag, 1982.
28. V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, 1998.