Complexity Measures of Supervised Classification Problems *

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

Mitra Basu
Dept of Electrical Engineering
City University of New York
140th Street & Convent Ave.,
New York, NY 10031, USA
basu@ccny.cuny.edu

September 1, 2001

Abstract

We studied a number of measures that characterize the difficulty of a classification problem, focusing on the geometrical complexity of the class boundary. We compared a set of real world problems to random labelings of points and found that real problems contain structures in this measurement space that are significantly different from the random sets. Distributions of problems in this space show that there exist at least two independent factors affecting a problem's difficulty. We suggest using this space to describe a classifier's domain of competence. This can guide static and dynamic selection of classifiers for specific problems as well as subproblems formed by confinement, projection, and transformations of the feature vectors.

Keywords: classification, clustering, complexity, linear separability, mixture identifiability

1 Introduction

Throughout the 1990's studies on multiple classifier systems have opened up many new opportunities for improving recognition accuracy [10][15]. But the empirically observed behavior of individual classifiers or combined systems is still strongly data dependent. We believe a better understanding of such data dependency is critical for further advances.

Such an understanding is not available in traditional theoretical studies that attempt to analyze classifier behavior for all possible problems and result inevitably in very weak performance bounds. Nor is it available from typical empirical studies that conclude with a presentation of accuracies on a small selection of problems, with little analysis on the reasons behind the classifier's success or failure. Comparative analysis of classifiers that relate their performances to data characteristics has received attention only recently [21][23]. In the project StatLog [21] a meta-learning attempt was made to predict the applicability of a classifier based on certain data characteristics. Such work focuses on the behavior of specific classifiers, and the reliability of the predictions is questionable. Typically the data measures are limited to statistical or information theoretic descriptions, but in classification it is the geometry that counts most.

Our study differs from such prior works in two critical aspects. First our emphasis is in the geometrical characteristics of the class distributions. We choose measures that can highlight the

*To appear, IEEE Transactions on Pattern Analysis and Machine Intelligence, March 2002.

manner in which classes are separated or interleaved, a factor most critical for classification accuracy. The measures are expected to affect performance of many classifiers that depend on the same kind of geometrical models. Second, previous studies use only a small number of problems (about 20 typically), and there is no additional knowledge about expected accuracies for each problem other than those obtained from the tested classifiers. Thus the study is somewhat circular. We overcome this by introducing synthetic datasets which by construction carry a certain accuracy expectation. Moreover, we use a well established procedure to identify problems that are linearly separable, i.e., considered easiest in pattern recognition. Thus we are able to show the goodness of the measures themselves with respect to these extreme cases. The scope of our study covers over 1,000 two-class problems, thus we have a much larger sample of cases.

2 The nature of classification difficulty

In reality, most practical classification problems arise from nonchaotic processes, many of which can be described by an underlying physical or behavioral model. Though the models may contain a stochastic component, there should still exist certain significant structure in the resulting class distributions that distinguishes them from random labelings. We believe that an analysis of such differences will provide us with a framework in which one can study the behavior of specific classifiers.

Structured data differ from random labeling in the difficulty of training a classifier to assign correct classes to future data from the same source. A random labeling is difficult since not much can be learned from the training data about the unseen points. On the other hand, in practical problems such learning can usually be done with various degrees of difficulty. In this paper we attempt to find a way to characterize this difficulty.

A problem can be difficult for different reasons. Certain problems are known to have nonzero Bayes error [11]. There the classes are ambiguous either intrinsically or due to inadequate feature measurements. This can be true regardless of sample size or feature space dimensionality. Others may have a complex decision boundary and/or subclass structures, so that no compact description of the boundary is possible. Again, this is independent of sample size and feature space dimensionality. Finally, small sample size and dimensionality induced sparsity introduce another layer of difficulty through a lack of constraints on the generalization rules.

In real world situations, often a problem becomes difficult because of a mixture of these three 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.

Class ambiguity is a property of the specific problem and chosen features, and is generally irrecoverable once the samples are taken and the features are computed. Thus our study focuses on the difficulty of a discrimination problem caused by the complexity of the decision boundary minimizing Bayes error. We will refer to the simplest (of minimum measure in the feature space) of such boundaries as the *class boundary*. With a complete sample, the class boundary can be characterized by its Kolmogorov complexity [17] [19], or the minimum length of a computer program needed to reproduce it (related concepts are also discussed in [5] [8]). 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. However, Kolmogorov complexity is known to be algorithmically uncomputable [20]. Thus we resort to relative measures that depend on the chosen descriptors. Specifically, we choose a number of geometrical descriptors that we believe to be relevant in the context of classi-

fication. The focus of this paper is on effective ways for characterizing the *geometrical complexity* of classification problems.
We assume that each problem is represented by a fixed set of training data consisting of vectors in $\mathbf{R}^d$ each associated with a class label. Furthermore, we assume that we have a sparse sample, i.e., there are unseen points from the same source that follow the same distribution but are unavailable during classifier design. The finite and sparse samples limit our knowledge about the boundary complexity, thus we are addressing only the *apparent* geometrical complexity of a problem based on a given training set.
In this study we discuss only two-class problems, because most of the measures we use are defined only for two-class discrimination. An n-class problem produces a matrix of two-class values for each chosen measure. To describe n-class problems, one needs a way to summarize such matrices. There are many possible ways to do so, especially if cost matrices are involved. This is a nontrivial problem that should be pursued after the measures are selected.

3 Measures of problem complexity

One practical measure of problem difficulty is the error rate of a chosen classifier. However, since our eventual goal is to study behavior of classifiers, we want to find other measures that are independent of such choices. Early explorations led us to the idea that a single descriptor may not suffice. Instead, we need to consider a number of different descriptors. In essence, we want to choose a feature space in which each classification problem can be represented as a point. Our study explores the distribution of real world problems in this space. If the points distribute in a meaningful continuum, then a problem's difficulty can be described by its position in this continuum. The same space may also be used to describe a classifier's domain of competence.
We investigate a number of measures from the literature of both supervised and unsupervised learning, as we believe that cluster structures can be essential characteristics for a discrimination problem. A few other measures are defined by us. All these measures are normalized as far as possible for comparability across problems. The measures we investigated can be divided into several categories.

3.1 Measures of overlap of individual feature values

3.1.1 Fisher's discriminant ratio (F1)

Fisher's discriminant ratio is a classic in this category:

$$f \;=\; \frac{(\mu_1 - \mu_2)^2}{{\sigma_1}^2 + {\sigma_2}^2}$$

where μ_1, μ_2, ${\sigma_1}^2$, ${\sigma_2}^2$ are the means and variances of the two classes respectively.
f as defined above is specific to one feature dimension. For a multidimensional problem, not necessarily all features have to contribute to class discrimination. As long as there exists one discriminating feature, the problem is easy. Therefore we use the maximum f over all the feature dimensions to describe a problem. The measure is referred to as F1.

3.1.2 Volume of overlap region (F2)

A similar measure is the overlap of the tails of the two class-conditional distributions. We can measure this by finding, for each feature, the maximum and the minimum values of each class, and

then calculating the length of the overlap region normalized by the the range of values spanned by both classes. We multiply the ratio thus obtained from each feature dimension to obtain a measure of the volume of the overlap region (normalized by the size of the feature space). Formally, let the maximum and minimum values of each feature f_i in class c_j be $max(f_i, c_j)$ and $min(f_i, c_j)$, then the overlap measure F2 is defined to be

$$\text{F2} = \prod_i \frac{MIN(max(f_i, c_1), max(f_i, c_2)) - MAX(min(f_i, c_1), min(f_i, c_2))}{MAX(max(f_i, c_1), max(f_i, c_2)) - MIN(min(f_i, c_1), min(f_i, c_2))}$$

where $i = 1, ..., d$ for a d-dimensional problem. Note that the volume is zero as long as there is at least one dimension in which value ranges of the two classes do not overlap.

3.1.3 Feature efficiency (F3)

With a high dimensional problem, we are concerned about how the discriminatory information is distributed across the features. Here we consider a measure of efficiency of individual features that describe how much each feature contributes to the separation of the two classes [12].
We consider a local continuity heuristic which assumes, for each feature, all points of the same class have values falling in between the maximum and minimum of that class. If there is an overlap in the feature values of two classes, we consider the classes ambiguous in that region along that dimension. Given that, a problem is easy (i.e. linearly separable) if there exists one feature dimension where the ranges of values spanned by each class do not overlap. For other problems that are globally unambiguous, one may progressively remove the ambiguity between the two classes by separating only those points that lie outside the overlapping region in each chosen dimension [12]. In each pass the features can be ordered by how many remaining points there are in the respective nonoverlapping regions. The *efficiency* of each feature is defined as the fraction of all remaining points separable by that feature. To represent the contribution of the feature most useful in this sense, we use the *maximum feature efficiency* (calculated with the entire training set) as a measure (F3).
In this procedure we consider only separating hyperplanes perpendicular to the feature axes. Therefore, even for a linearly separable problem, F3 may be less than 1 if the optimal separating hyperplane happens to be oblique. In other words, the joint effects of the features are not accounted for by this measure.

3.2 Measures of separability of classes

3.2.1 Linear separability (L1, L2)

Linear separability was intensively discussed in the early literature. Many algorithms were proposed to determine linear separability, most of which can only arrive at positive conclusions and may iterate indefinitely for negative cases. In a recent study we found that linear programming methods far outperform the adaptive methods in terms of definiteness and correctness of decisions and time efficiency [1]. To handle both separable and nonseparable cases, we use a formulation proposed by Smith [24] that minimizes an error function:

$$\begin{array}{ll} \text{minimize} & \mathbf{a}^t\mathbf{t} \\ \text{subject to} & \mathbf{Z}^t\mathbf{w} + \mathbf{t} \geq \mathbf{b} \\ & \mathbf{t} \geq \mathbf{0} \end{array}$$

where **a**, **b** are arbitrary constant vectors (both chosen to be **1**), **w** is the weight vector, **t** is an error vector, and **Z** is a matrix where each column **z** is defined on an input vector **x** (augmented by adding one dimension with a constant value 1) and its class c (with value c_1 or c_2) as follows:

$$\begin{cases} \mathbf{z} = +\mathbf{x} & \text{if } c = c_1 \\ \mathbf{z} = -\mathbf{x} & \text{if } c = c_2. \end{cases}$$

The value of the objective function in this formulation is used as a measure (L1). It is zero for a linearly separable problem. Notice that this measure can be heavily affected by outliers that happen to be on the wrong side of the optimal hyperplane. We normalize this measure by the number of points in the problem and also by the length of the diagonal of the hyperrectangular region enclosing all such points in the feature space. We also include the error rate of such a linear classifier on the original training set as a measure (L2).

3.2.2 Mixture identifiability (N1, N2, N3)

Friedman and Rafsky [7][25] proposed a test on whether two samples are from the same distribution. It is thus useful for deciding if the points labeled as two classes form separable distributions. The method relies on computing a minimum spanning tree (MST) that connects all the points to their nearest neighbors (regardless of class). Then the number of points connected to the opposite class by an edge in this MST are counted. These are considered to be the points lying next to the class boundary (Figure 1). The fraction of such points over all points in the dataset is used as a measure (N1).

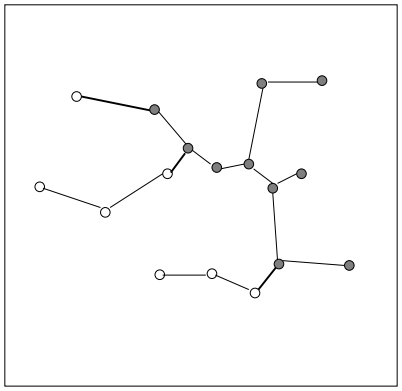


Figure 1: A minimum spanning tree connecting points of two classes. The thicker edges connect two classes.

Understandably for heavily interleaved or randomly labeled data, a majority of points will appear next to the class boundary. However, the same can be true for a linearly separable problem with margins narrower than the distance between points of the same class.
A closely related measure is defined as follows. We first compute the Euclidean distance from each point to its nearest neighbor within or outside the class. We then take the average of all the distances to intra-class nearest neighbors, and the average of all the distances to inter-class nearest

neighbors. The ratio of the two averages is used as a measure (N2). This measure compares the dispersion within the classes to the gap between the classes. While the MST based measure is sensitive to which (intra or inter class) neighbor is closer to a point, this measure takes into account the magnitudes of the differences.
The proximity of points in opposite classes obviously affects the error rate of a nearest neighbor classifier. Thus we include a leave-one-out estimate of the nearest neighbor error rate as another measure (N3).

3.3 Measures of geometry, topology, and density of manifolds

Some measures are intended to describe the geometry of the manifolds spanned by each class. These include various estimators of intrinsic dimensionality of the dataset. Others attempt to describe the shapes of the manifolds, the existence of isolated submanifolds, or variation in the point densities within the manifolds, such as methods and tests suggested in [4] [27] [28] for the data distribution against hypotheses of uniformity or normality. We investigate two measures of this category.

3.3.1 Nonlinearity (L3, N4)

Hoekstra and Duin [14] proposed a measure for the *nonlinearity* of a classifier w.r.t. to a given dataset. Given a training set, the method first creates a test set by linear interpolation (with random coefficients) between randomly drawn pairs of points from the same class. Then the error rate of the classifier (trained by the same training set) on this test set is measured. This measure is sensitive to the smoothness of the classifier's decision boundary as well as the intrusion of the convex hull of each class into that of the other. We consider the nonlinearity of the linear classifier defined in Section 3.2.1 (L3) and that of a nearest neighbor classifier (N4).

3.3.2 Space covering by ϵ-neighborhoods (T1)

The local clustering properties of a point set can be described by an ϵ-neighborhood pretopology [18]. Here we consider a reflexive and symmetric binary relation $\mathcal{R}$ of two points x and y in a set F. $\mathcal{R}$ is defined by $x\mathcal{R}y \Leftrightarrow d(x,y) < \epsilon$, where $d(x,y)$ is a given metric and ϵ is a given nonzero constant. Define $\Gamma(x) = \{y \in F | y\mathcal{R}x\}$, an adherence mapping ad from the power set $\mathcal{P}(F)$ to $\mathcal{P}(F)$ is such that

$$\begin{cases} ad(\phi) = & \phi \\ ad(\{x\}) = & \{x\} \cup \Gamma(x) \\ ad(A) = & \bigcup_{x \in A} ad(\{x\}) \quad \forall A \subset F. \end{cases}$$

Adherence subsets can be grown from a singleton $\{x\}$: $\{x\} = ad^0(\{x\})$, $ad(\{x\}) = ad^1(\{x\})$, ..., $ad(ad^n(\{x\})) = ad^{n+1}(\{x\})$, where j is called the *adhesion order* in $ad^j(\{x\})$. From a point of each class one can grow successive adherence subsets to the highest order n such that $ad^n(\{x\})$ includes only points of the same class but $ad^{n+1}(\{x\})$ includes points of the opposite class.
To represent a class, we eliminate any adherence subsets that are strictly included in another one. For each point, only the highest order n adherence subset is kept such that all elements of $ad^n(\{x\})$ are within the class of x. Using the ϵ-neighborhoods with Euclidean distance as $\mathcal{R}$, each retained adherence subset associated with a point is the largest hypersphere that contains it and no points from the other class, in units of the chosen ϵ (Figure 2). We use a value of ϵ that is 0.55δ where δ is the distance between two closest points of opposite classes. The value is chosen such that the lowest adhesion order is always zero, occurring at the points lying closest to the class boundary.

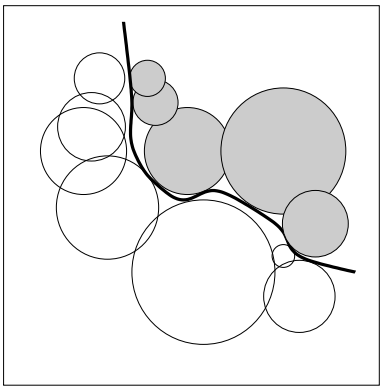


Figure 2: Retained adherence subsets for two classes near the boundary.

A list of such ϵ-neighborhoods needed to cover the two classes is a composite description of the shape of the classes. This is an interior description rather than a boundary description as given by the MST based measures. The number and order of the retained adherence subsets indicate how much the points tend to be clustered in hyperspheres or distributed in thinner structures. In a problem where each point is closer to points of the other class than points of its own class, each adherence subset is retained and is of a low order. We normalize the count of the retained adherence subsets by the total number of points. The normalized count is referred to as measure T1.

Others (T2)

The relevance of other measures is less obvious. For example, it is not clear what can be inferred from the intrinsic dimensionality of a problem without differentiation by class. A problem can be very complex even if embedded in a low dimensional space (e.g. randomly labeled points along a one-dimensional space have a complex class boundary). Also, variation in density within a manifold seems irrelevant as long as the manifolds can be easily separated. Similarly, existences of submanifolds of one class surrounding the other (e.g. consider two classes black and white on a checkerboard) may make a problem difficult for, say, a linear classifier, but may not affect a nearest neighbor classifier by much. Nevertheless, in discussions on *curse of dimensionality*, the number of samples is often compared to the number of feature dimensions. To relate to such discussions, we include the average number of samples per dimension as another measure (T2).

4 Experimental Setup

Table 1 summarizes the measures we include in the study. These measures define a 12-dimensional measurement space, and every classification problem is represented by a point in this space. We study the distribution of some real world problems in this space as well as some artificial problems where we have control over data generation.

We consider two collections of problems. The first collection comes from the UC-Irvine Machine Learning Depository [3]. We selected 14 of the datasets that contain at least 500 points and no missing values: *abalone, car, german, kr-vs-kp, letter, lrs, nursery, pima, segmentation, splice, tic-tac-toe, vehicle, wdbc, and yeast.* The problems we consider are discrimination between all pairs of classes in these 14 data sets. For those sets containing categorical features, the values are numerically coded. In total there are 844 two-class discrimination problems. These problems

F1	maximum Fisher's discriminant ratio
F2	volume of overlap region
F3	maximum (individual) feature efficiency
L1	minimized error by linear programming (LP)
L2	error rate of linear classifier by LP
L3	nonlinearity of linear classifier by LP
N1	fraction of points on boundary (MST method)
N2	ratio of average intra/inter class NN distance
N3	error rate of 1NN classifier
N4	nonlinearity of 1NN classifier
T1	fraction of points with associated adherence subsets retained
T2	average number of points per dimension

Table 1: List of investigated measures.

originated from a variety of physical and behavioral processes.
The second set consists of 100 artificial two-class problems each having 1000 points per class. Problem 1 has one feature dimension, problem 2 has two, so forth and the last problem contains 100 features. Each feature is a uniformly distributed pseudorandom number in $[0, 1]$. The points are randomly labeled as one of the two classes. Therefore, these are intrinsically complex problems, and they should delimit one end of an complexity spectrum. We created these for comparison and contrast with real world data, and will refer to them as the random noise sets.

5 Results and Discussions

We implemented algorithms for calculating each measure and applied them to each of the 944 problems (844 real and 100 artificial). We then examined the distribution of these 944 points in this space by the density plots and pairwise scatter plots (log-scaled if necessary) for interesting structures. Of the 844 problems, 452 are found to be linearly separable by a linear programming procedure [1]. The class boundary (if only the training set is concerned) of these problems can be described by the coefficients of the separating hyperplane, so by Kolmogorov's notion these are simple problems. We thus expect these to delimit the other end of any complexity spectrum. In order to compare the distributions of these three types of problems (linearly separable, nonseparable, and random noise), we mark them differently in each plot. We also add a small constant to the zero values so that those points can be included in a logarithmic scale.

General observations

Figure 3 shows the distributions of the three types of problems over the values of each measure. From the univariate distributions, we observe that several measures (F1, F2, F3, L2, L3) are especially effective in assigning the linearly separable problems and random noise sets to opposite ends of the ranges. The measures N1,N2,N3 have this effect too except for a curious peak at the far right for some linearly separable problems. Close examination of the values reveals that those are extreme cases where there are only two or three points in the training set. These problems are linearly separable but the nearest neighbors are always in the wrong class. Measures L1,N4,T2 are uninteresting on their own since the distributions of the three types of problems overlap heavily. With measure T1, the three types of problems overlap but they have different dispersions. Random

noise sets are highly concentrated at an extreme and the easier problems are more dispersed. This may be useful for ruling out certain possibilities for a new problem.
None of the measures can cleanly separate all three types of problems. For example, with F1 the random noise sets overlap heavily with the linearly nonseparable problems, and with measures F2,F3,L2,L3,N1,N2,N3, no clear distinction exists between the linearly separable and nonseparable problems. However, given that the distributions overlap in different ways with different measures, it is possible that some separation can be seen in the bivariate or multivariate distributions.
Several bivariate plots show good separation between the three types of problems (Figure 4(a)) while some others show more blurred boundaries (4(b)). We also observe that the points span a fan-like structure in many plots, with random noise sets and the linearly separable problems clustered near opposite ends (e.g. Figure 5(a)). This leads us to believe that a problem's complexity can be decomposed into at least two independent aspects. In some other plots the distribution is close to one-dimensional, which means that the two measures are highly correlated (Figure 5(b)).
In several plots we see the random noise sets appearing on the boundary of the problem distribution, and they stay far from the real problems due to their exaggerated difficulty (e.g. Figure 6(a)). This suggests that real world problems often contain structures that are significantly different from random labelings. Interestingly, there exist large differences between various random noise sets. By examining their appearances in the plots involving other measures, we find that such differences are due to the apparent simplicity caused by sparsity of samples in the higher dimensional problems (Figure 6(b)). A simple classifier obtained with these training sets will turn out to perform very badly on unseen points from the same problem.

Groups of measures

The observed patterns in the 12-dimensional measurement space reveal the multi-faceted nature of a problem's complexity. Alternatively, this says that though most of our chosen measures may be biased towards some particular characteristics of a problem, not all of them are measuring the same thing. For example, consider the linear programming method for determining linear separability. The amount of error produced by the LP method (the minimized objective function, measure L1) is a measure of the nonlinearity of a given problem. However, we know that a problem that is highly nonlinear can sometimes be easily classified by a nearest-neighbor classifier. So, such measures are classifier dependent and thus cannot provide an *absolute* scale. In order to gain insight into such biases, we group the measures by similarity in their definitions:

- Linear classifier based - L1,L2,L3
- Nearest-neighbor based - N1,N2,N3,N4
- Geometry or topology based - F1,F2,F3,T1,T2

The first two groups are classifier dependent whereas the last one is not. In the next few paragraphs we analyze the information that can be extracted from the bivariate distributions that involve two measures of the same or different groups. We expect that combining measures from different groups delivers more information about a problem. It is our hope that a *judicious* combination of such measures may bring out most, if not all, aspects of the problem complexity.

Linearity versus proximity

Consider the measures in the first two groups. Measures of the same group are highly correlated (e.g. L2,L3 in Figure 5(b), N1,N3 in 7(a)) whereas measures of different groups are uncorrelated

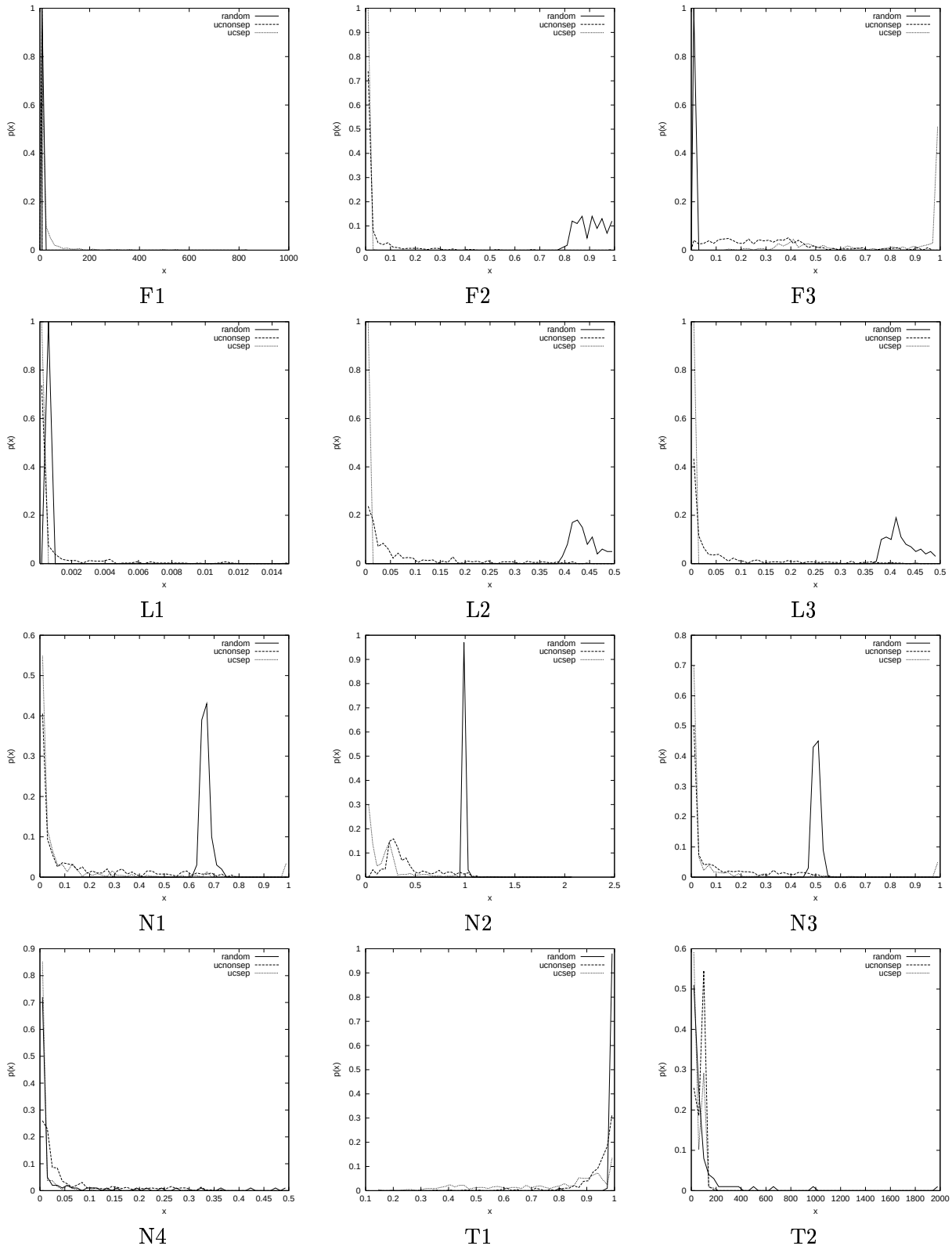
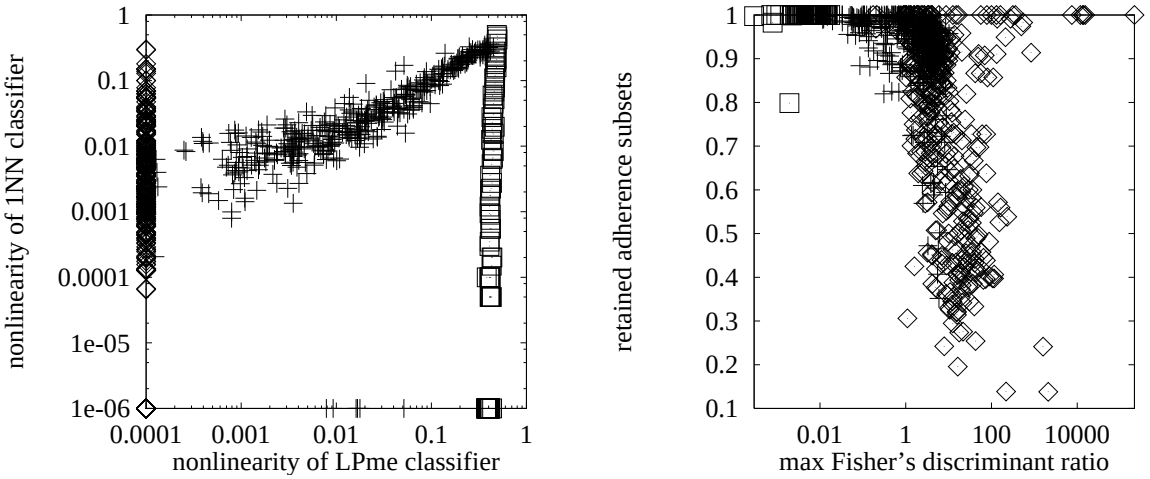
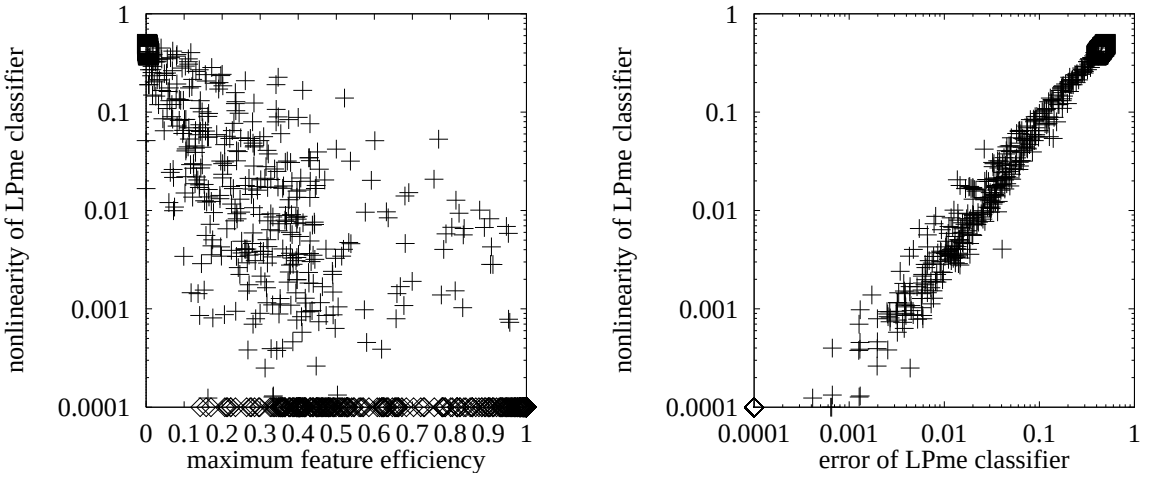


Figure 3: Distributions of the three types of problems in each measure.



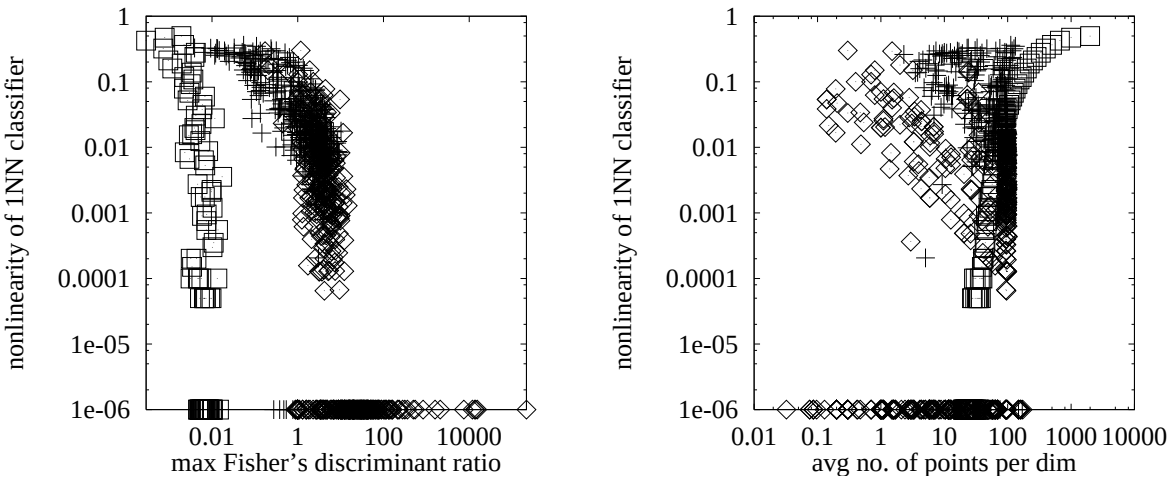
(a) L3,N4 (b) F1,T1

Figure 4: Clean (a) or blurred (b) separations between the three types of problems ($\diamond$: linearly separable problems; $+$: linearly nonseparable problems; $\square$: random noise sets).



(a) F3,L3 (b) L2,L3

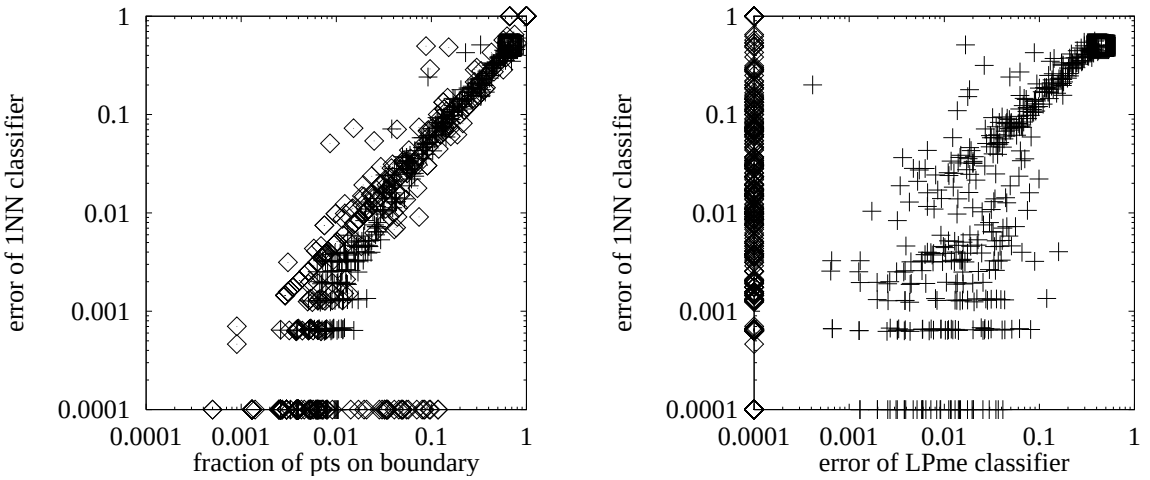
Figure 5: Pairs of measures that describe different (a) or similar (b) aspects of a problem.



(a) F1,N4 (b) T2,N4

Figure 6: Differences between random labelings and ordinary problems (a) and their correlation with sample sparsity (b).

(e.g. L3,N4 in Figure 4(a), L2,N3 in 7(b)). It can be observed from Figure 7(b) that the error rate of an NN classifier is unrelated to that of the LP based linear classifier. This is not totally unexpected since the two classifiers operate by different principles. The linear classifier is sensitive to the *location* of misplaced points (i.e., points that belong to one class appear inside the other class) whereas the NN classifier is sensitive to the *number* of misplaced points. That is, one is more geometry-sensitive and the other is density-sensitive. So, if a problem has high values for both measures, then one concludes that many samples from one class appear in *strategic* locations inside the other class. Such problems are indeed very complex, and we observe that the random noise sets have this property. With other problems, the uncorrelated error rates manifest the difficulty of defining problem complexity on accuracies of specific classifiers.



(a) N1,N3 (b) L2,N3

Figure 7: (a) Correlated measures of the same group. (b) Uncorrelated measures of different groups show better separation between the three types of problems.

Among the linearly separable problems, no ordering can be derived with any measure from the first group, though such measures are informative for other problems. However, measure L1 differs from others in that, the random noise sets are not at or close to the extreme values. So while it tells something about the existences and locations of outliers in a problem, and it correlates well with a few other measures (e.g. L2,L3,N1) for linearly nonseparable problems, its does not serve well as an independent complexity scale.

Geometry and topology

Measures in the third group describe the geometrical or topological properties of a problem without a specific classifier model. Consider the plots involving Fisher's discriminant ratio, F1. Nonlinear problems have lower value for this measure than the linear problems. When it is combined with the volume of overlap measure F2 (Figure 8(a)), linear problems tend to appear at the lower right corner and the nonlinear problems at the upper left hand corner. There is a region in the graph where both types of problems appear. A linear problem belonging to this region may have an oblique separating hyperplane so that projections of the points to the axes show nonzero overlap. A characteristic like this may have implications for certain classifier designs. We notice similar groupings when F1 is combined with N1,N2 and N3.

Measure F2, volume of overlap region, correlates well with F1 and F3 from the same group and jointly they provide better separation between the three types of problems roughly along the diagonal. (Figure 8(a)(b)). However, it does not describe the shape of the class boundary, as observed from plots F2,L3 and F2,T1 (Figure 8(c)(d)) where non-random problems with the same degree of class overlap can have any amount of nonlinearity or adherence subsets. However, problems that are closer to the random noise sets show smaller variation in these several measures. F3 is similar to F2 in such combinations.

Measure T1, as indicated earlier, is an interior description of the classes. With T1, random noise sets tend to concentrate on one end but no perceptible separation is observed between the linearly separable and nonseparable problems. The type-dependent variations in dispersion of the problem distributions create a fan-like structure when T1 is combined with other measures that separates

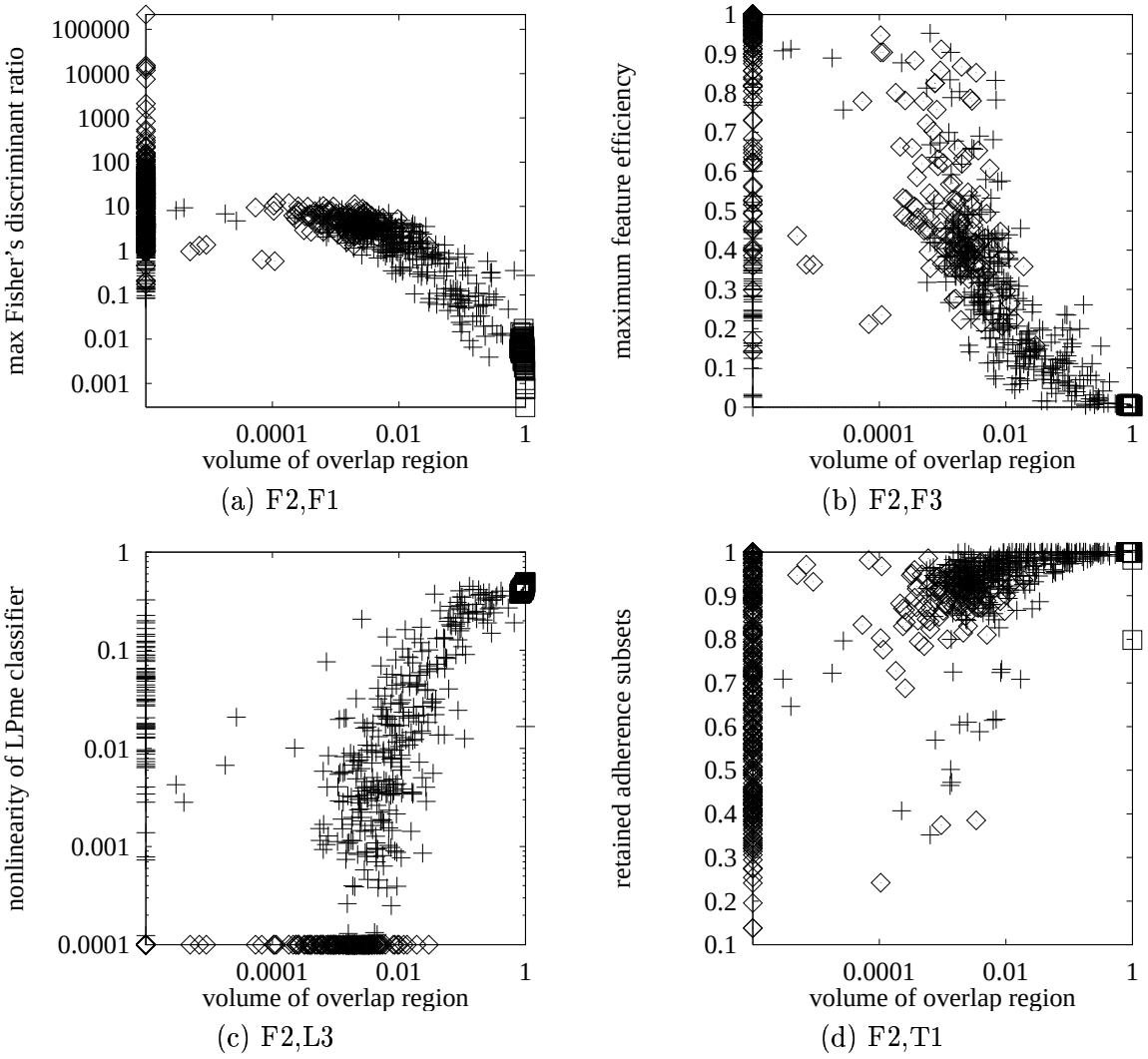
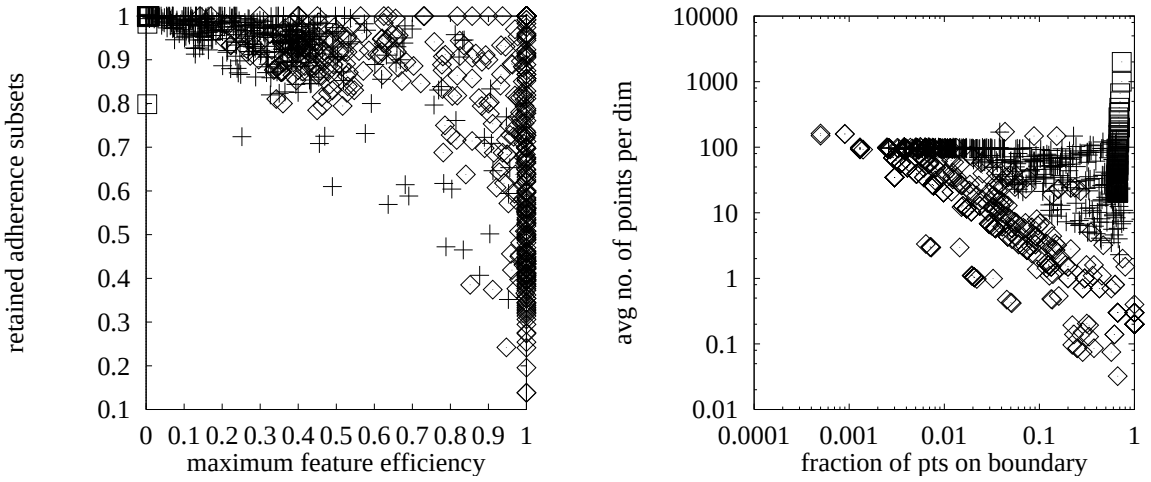


Figure 8: (a)(b) Joint usage of two measures shows different characteristics of class overlap. (c)(d) Non-random problems with the same degree of class overlap can have any amount of (c) nonlinearity or (d) adherence subsets. But the variation is much smaller as the problems approach the random noise sets.

the problem types (e.g. F3,T1 in Figure 9(a)). The plots involving this measure enforce our belief that problem complexity depends more on the shape of the class boundary than on the shape of the classes away from the boundary. Measure T2, average number of points per dimension, is useful only when combined with others. Note that it spreads out an order among the linearly separable problems when combined with measures N1, N2, N3, N4 (Figure 9(b), see also Figure 6(b)).



(a) F3,T1 (b) N1,T2

Figure 9: (a) The fan structure caused by differences in dispersion of the three types of problems and their separation. (b) Joint usage of T2 and others yields better separation of the three types of problems.

6 Principal Component Analysis

As discussed above, some measures are highly correlated. A challenge for the future is to determine the intrinsic dimensionality of the problem distributions in this space, and identify the independent factors. As a first step, we perform a principal component analysis on the distribution of the 944 problems in the 12-dimensional space. To bring the values to a comparable scale, we first standardize each measure by subtracting the mean and dividing by the standard deviation. Then the principal components are estimated using the S-plus software.

The results show that there are six components each explaining more than 5% of the variance (Table 2). The first component, explaining over 50% of the variance, has almost even contributions from measures F2,L2,L3,N1,N2,N3. It describes the joint effect of linearity of class boundaries and proximity of opposite class neighbors. Measures T1,T2 contribute strongly (and negatively) to the second component but their effects are opposite to those of F1,F3,N1 and N3 which have positive weights. This appears to be the balance between effects of (1) sampling density and within-class scatter and (2) between-class proximity. Measures L1,N4 dominate the third component and both are measures of intrusion of samples into the wrong class. Their effects are offset by F2 which is also a measure of overlap but is calculated from ranges of projections of points onto the feature axes. Thus this component appears to relate to the concentration and orientation of the intrusion in the

feature space. F1 and T1 dominate the fourth component and both are affected by within-class scatter. The projections of the problem distribution onto the first four components are shown in Figure 10(a)(b). The separations between the three types of problems are visible in Figure 10(a) but are more obvious in Figure 10(b).

Component	C1	C2	C3	C4	C5	C6
Prop. of Var.	0.5033	0.1162	0.1064	0.0859	0.0761	0.0521
Cum. Prop.	0.5033	0.6195	0.7259	0.8118	0.8879	0.9400
Loadings						
F1	0.01	0.26	0.03	0.86	-0.26	-0.33
F2	0.33	0.08	-0.43	-0.12	-0.09	-0.20
F3	-0.29	0.42	0.03	-0.11	-0.32	0.29
L1	0.17	0.08	0.68	-0.15	0.00	-0.36
L2	0.38	0.04	-0.15	-0.14	-0.10	-0.24
L3	0.38	0.05	-0.16	-0.14	-0.12	-0.23
N1	0.36	0.30	0.04	0.01	-0.04	0.36
N2	0.37	-0.02	0.02	0.03	0.12	-0.03
N3	0.32	0.36	0.00	0.11	-0.03	0.49
N4	0.24	-0.20	0.52	-0.04	-0.35	0.16
T1	0.23	-0.32	0.07	0.37	0.57	0.28
T2	0.08	-0.61	-0.15	0.13	-0.58	0.22

Table 2: Proportion and cumulative proportion of variance explained by the first six principal components and their loadings on the measures. The loadings are the coefficients of the projection associated with each p.c.

7 Case Studies

How does a new problem compare to those appearing in the study? We select two sets of problems to illustrate this use of the measures. The first set consists of the three two-class problems in the well studied Iris dataset [6]. The dataset contains three classes {Iris-setosa, Iris-versicolor, Iris-virginica} among which only the pair {Iris-versicolor, Iris-virginica} is linearly nonseparable. Each class contains 50 points in a 4-dimensional space. We compute the complexity measures for all the three problems and show their positions in the principal component projections in Figure 11(a)(b) which have the same scales as Figure 10(a)(b). Different symbols are used for the linearly separable and nonseparable cases to highlight their difference.

The second set consists of 100 two-class problems. Each problem has two classes, 100 points per class, and resides in a one-dimensional feature space. The points are normally distributed around mean $+m_i$ for the positive class and $-m_i$ for the negative class, where $m_i = i$ for the ith problem ($i = 0, ...99$), with standard error e fixed at 30. Thus the two classes in the first problem overlap completely and separability increases progressively for the other problems. We compute the complexity measures to see the trajectory these problems follow in the measurement space. Again we show their positions in the principal component projection (Figure 11(a)(b)). Notice that the 0th problem (with $m = 0$) is not included in the plots because of an infinity in the averaged intra/inter class distances. The next most difficult problem (with $m = 1$) is highlighted with a different symbol. An interesting observation is that these problems follow a different "trajectory

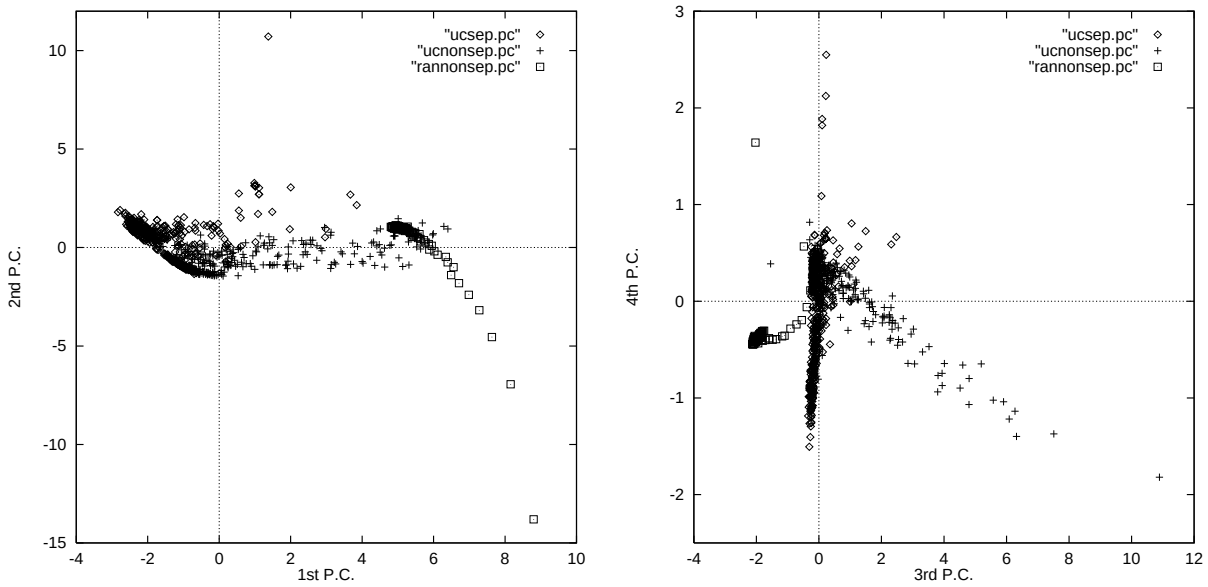


Figure 10: Projections of problem distributions onto the (a) 1st and 2nd and (b) the 3rd and 4th principal components.

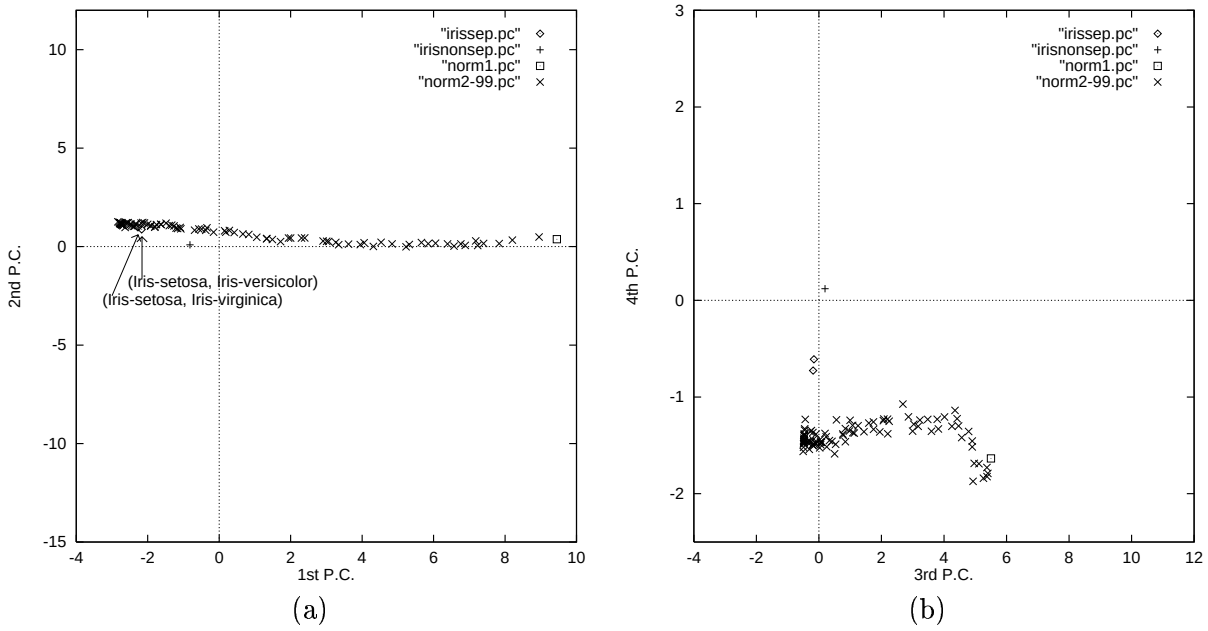


Figure 11: Projections of the new cases onto the principal components found with the 944-problem collection.

of difficulty" from that of the random noise sets, since they include both linearly separable and nonseparable cases.
In practice, comparison of problems in this way can lead to more realistic expectations of classification accuracy. There are also other uses of these measures. Already we see that different classifiers (say, linear classifiers and nearest neighbors) have different domains of competence, and studies of other classifiers' domains in this space would be interesting. For instance, in [9] a study is reported where the measures are related to the comparative advantages of two methods to construct decision forests. Moreover, it is conceivable that new algorithms may be constructed using results of this study. Systematic procedures may be found to transform a given problem to a space where the problem complexity (as characterized in this study) is reduced. Certain existing methods, such as stochastic discrimination [16] and support vector machines [26], pursue this implicitly. It is hoped that our study will encourage studies of more explicit forms of such procedures.

8 Conclusions

We studied several measures for characterizing the complexity of classification problems. We found that there exist rich structures in such a measurement space that reveal the intricate relationships among the factors affecting the difficulty of a problem. The distribution of real world problems is significantly different from that of random noise, which suggests that many classification tasks arising naturally from real-life processes do contain learnable structures.
Here we examined the structure of only the given training set of a problem. Difficulty of real problems also lies in generalizing the classification to unseen points. To what extent a training set represents a test set should be discussed in the context of generalization ability of classifiers. For this we refer readers to Kleinberg's arguments on M-representativeness [16], Berlind's hierarchy of indiscernibility [2], Vapnik's VC-dimension theory and his analysis on small sample effects [26], and observations and discussions about several classifiers by Raudys and Jain [22]. An interesting question for further investigation is the consistency of our chosen measures on bootstrap samples of the training set.
It should be noted that our analysis of the training data can also be applied to their subsets or transformations, that is, data confined in selected regions, projected onto selected subspaces, or transformed to another space. Corresponding choices of classifiers can be made for these altered datasets as well. This can lead to a way of designing static or dynamic classifier selection schemes, e.g., to choose different classifiers for data falling into different branches of a decision tree.
Finally, we have looked into only a tiny set of problems among all those possible, and any extrapolation of results must be done with extreme caution. Many open questions remain to be answered. Will the empty regions in the chosen measurement space be filled with some new problems? Or do they represent some constraints of geometry and topology? How should sampling density be involved, so that proper qualifications can be made on characterizations of extreme cases like two-point training sets? Nevertheless, we hope this study will open a potentially fruitful path into a better understanding of classification problems and classifier behavior.

Acknowledgements

The authors would like to thank the AE and the reviewers for many suggestions on the manuscript.

References

[1] M. Basu, T.K. Ho, The learning behavior of single neuron classifiers on linearly separable or nonseparable input, *Proceedings of the 1999 International Joint Conference on Neural Networks*, Washington, DC, July 1999.

[2] 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.

[3] C.L. Blake, C.J. Merz, UCI Repository of machine learning databases, http://www.ics.uci.edu/ mlearn/MLRepository.html, University of California, Department of Information and Computer Science, Irvine, CA, 1998.

[4] D.S. Broomhead, R. Jones, G.P. King, Topological dimension and local coordinates, *Journal of Physics, A: Math. Gen.*, **20**, 6, 1987, L563-L569.

[5] G.J. Chaitin, A theory of program size formally identical to information theory, *Journal of the ACM*, **22**, 1975, 329-340.

[6] R.A. Fisher, The use of multiple measurements in taxonomic problems, *Annual Eugenics*, **7**, Part II, 1936, 179-188.

[7] J.H. Friedman, L.C. Rafsky, Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests, *The Annals of Statistics*, **7**, 4, 1979, 697-717.

[8] M. Gell-Mann, What is complexity? *Complexity*, **1**, 1, 1995, 16-19.

[9] T.K. Ho, Complexity of classification problems and comparative advantages of combined classifiers,, *Proceedings of the 1st International Workshop on Multiple Classifier Systems*, Cagliari, Italy, June 21-23, 2000, 97-106.

[10] T.K. Ho, Multiple classifier combination: Lessons and next steps, in A. Kandel, H. Bunke, (eds.), *Hybrid Methods in Pattern Recognition*, World Scientific, 2002.

[11] T.K. Ho, H.S. Baird, Large-scale simulation studies in image pattern recognition, *IEEE Trans. on PAMI*, **19**, 10, October 1997, 1067-1079.

[12] T.K. Ho, H.S. Baird, Pattern classification with compact distribution maps, *Computer Vision and Image Understanding*, **70**, 1, April 1998, 101-110.

[13] T.K. Ho, M. Basu, Measuring the complexity of classification problems, *Proceedings of the 15th ICPR*, Barcelona, Spain, September 3-8, 2000, **2**, 43-47.

[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 Trans. on PAMI*, **22**, 1, January 2000, 4-37.

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

[17] A.N. Kolmogorov, Three approaches to the quantitative definition of information, *Problems of Information Transmission*, **1**, 1965, 4-7.

[18] F. Lebourgeois, H. Emptoz, Pretopological approach for supervised learning, *Proceedings of the 13th International Conference on Pattern Recognition*, Vienna, 1996, 256-260.

[19] M. Li, P. Vitanyi, *An Introduction to Kolmogorov Complexity and Its Applications*, Springer-Verlag, 1993.

[20] J.M. Maciejowski, Model discrimination using an algorithmic information criterion, *Automatica*, **15**, 1979, 579-593.

[21] D. Michie, D.J. Spiegelhalter, C.C. Taylor (eds.), *Machine Learning, Neural and Statistical Classification*, Ellis Horwood, 1994.

[22] S. Raudys, A.K. Jain, Small sample size effects in statistical pattern recognition: Recommendations for practitioners, *IEEE Trans. PAMI*, **13**, 3, 1991, 252-264.

[23] S.Y. Sohn, Meta analysis of classification algorithms for pattern recognition, *IEEE Trans. PAMI*, **21**, 11, 1999, 1137-1144.

[24] F.W. Smith, Pattern classifier design by linear programming, *IEEE Transactions on Computers*, **C-17**, 4, April 1968, 367-372.

[25] S.P. Smith, A.K. Jain, A test to determine the multivariate normality of a data set, *IEEE Trans. PAMI*, **10**, 5, Sep. 1988, 757-761.

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

[27] P.J. Verveer, R.P.W. Duin, An evaluation of intrinsic dimensionality estimators, *IEEE Trans. PAMI*, **17**, 1, Jan. 1995, 81-86.

[28] N. Wyse, R. Dubes, A.K. Jain, A critical evaluation of intrinsic dimensionality algorithms, *Pattern Recognition in Practice*, E.S. Gelsema and L.N. Kanal (eds.), North-Holland, 1980, 415-425.