Measuring the Complexity of Classification Problems

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

Mitra Basu
Dept of EE, City University of New York
140th Street & Convent Ave.,
New York, NY 10031, USA
eemb@ee-mail.engr.ccny.cuny.edu

Abstract

We studied a number of measures that characterize the difficulty of a classification problem. We compared a set of real world problems to random combinations of points in this measurement space and found that real problems contain structures that are significantly different from the random sets. Distribution of problems in this space reveals that there exist at least two independent factors affecting a problem's difficulty, and that they have notable joint effects. 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, projections, and transformations of the feature vectors.

1. Introduction

Many theoretical studies in pattern recognition attempt to analyze the behavior of classifiers for all possible problems, i.e., classes defined on arbitrary combinations of points in a feature space. On the other hand, empirical studies often conclude with a presentation of the error rates of a classifier on a small selection of real problems, with little analysis on the reasons behind the classifier's success or failure. Comparative analysis of classifiers and how their performances relate to data characteristics has received attention only very recently [9].

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 data distributions that differs from a random combination of points. We believe that an analysis of such differences will provide us with a framework for studying classifier behavior.

Structured data differ from random combinations in the difficulty of obtaining a classifier that can assign correct class labels for data from the same source. Points with randomly assigned labels are difficult since not much can be learned from the training data about the unseen points. With real world recognition data such learning can often be done with various degree 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 [4]. Others may have a complex decision boundary and/or subclass structures. Sometimes high dimensionality of the feature space and sparseness of available samples lead to estimation difficulties.

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 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 will consider a number of different descriptors. In essence, we are choosing a feature space in which each classification problem can be represented as a point. We are interested in the distribution of selected real world problems in this space. We attempt to determine if there exists any continuum, such that a problem's difficulty can be described by its position in this continuum. We also conjecture that the same space can be used to describe a classifier's domain of competence.

We assume each problem is represented by a fixed set of training data consisting of a collection of vectors in $\mathbf{R}^n$ each associated with a class label. In this study we discuss only two-class problems. Furthermore, we assume that we have a sparse sample, not all possible points from the same source are available for classifier design.

2. Measures of problem complexity

The complexity of a discrimination problem is the complexity of its decision boundary that minimizes Bayes error. We will refer to the simplest (of minimum measure in the input space) of such boundaries as the *class boundary*. With a complete sample, the class boundary can be characterized by its Kolmogorov complexity [7], or the minimum length of a computer program needed to reproduce it. A problem

is difficult 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.

An incomplete or sparse sample adds another layer of complexity 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.

We investigated a number of measures previously proposed in the literature to describe classification problems. We borrowed from the studies of both supervised learning and unsupervised learning, as we believe that cluster structures can also be essential characteristics for a discrimination problem. A few other measures are defined by ourselves. All these measures are normalized as far as possible for comparability across problems. The measures we examined can be divided into several categories.

2.1. Measures of overlap of individual feature values

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 highly discriminating feature, the problem can be easy. Therefore we use the maximum f over all the feature dimensions to describe a problem.

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 from which we calculate 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). Note that the volume is zero as long as there is at least one dimension in which the two classes do not overlap.

2.2. Measures of separability of classes

Linear separability 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, for determining linear separability, 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 [10] 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 $\mathbf{a}$, $\mathbf{b}$ are arbitrary constant vectors (both chosen to be $\mathbf{1}$), $\mathbf{w}$ is the weight vector, $\mathbf{t}$ is an error vector, and $\mathbf{Z}$ is a matrix where each column $\mathbf{z}$ is defined on an input vector $\mathbf{x}$ (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. It is zero for a linearly separable problem. 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 the samples in the feature space.

Mixture identifiability Friedman and Rafsky [3][11] 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. The fraction of such points over all points in the dataset is used as a measure.

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 a margin narrower than the distance between points of the same class.

A closely related measure is defined as follows. We first compute the distances 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. We use the ratio of the two averages as a measure. 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.

2.3. Measures of geometry, topology, and internal 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. 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 tests suggested in [13] [14] for data distributions against hypotheses of uniformity or normality. We investigated two measures of this category.

Hoekstra and Duin [5] 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 given 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 overlap of the convex hulls of the two classes. We consider the nonlinearity of a linear classifier (minimizing error) and that of a nearest neighbor classifier. We also include the error rate of that linear classifier on the original training set and the error of the nearest neighbor classifier estimated by leave-one-out.

In addition, we considered a supervised k-means clustering procedure. In this procedure, we first calculate the centroid of each class. Then we remove all points lying closer to the centroid of its own class than that of the other class. Next, we update the centroids for the remaining points and repeat the procedure, until either no more points can be removed or one of the class has no points left. The number of iterations it takes is used as a measure. This measure is sensitive to the difference in variances of the two classes, and also to the amount of overlap of the convex hulls of the two classes.

The relevance of other measures are less obvious. It is not clear what role is played by the intrinsic dimensionality of a problem without differentiation by class. A problem can be very complex even if embedded in a low dimensional space (we will show an artificial example). Also, variation in density within a manifold seems irrelevant as long as the manifolds can be easily separated. Similarly, existence of submanifolds of one class surrounding the other may make a problem difficult for, say, a linear classifier, but may not affect a nearest neighbor classifier by much.

3. Sources of Data

We considered two collections of problems. The first collection is from the UC-Irvine Machine Learning Depository. 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 considered are discrimination between all pairs of classes in these 14 data sets. Categorical features were numerically coded. Totally there are 844 two-class discrimination problems. These problems originated from a variety of physical and behavioral processes.

The second set consists of 100 artificial two-class problems each has 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 any spectrum of difficulty. We created these for comparison and contrast with real world data, and will refer to them as the random noise sets.

4. Results and Discussions

Table 1 summarizes the measures we included in the study. 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 all the pairwise scatter plots (two-dimensional projections) for interesting structures. Of the 844 problems, 452 are found to be linearly separable by a linear programming procedure [1]. 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 difficulty spectrum. In order to compare the distributions of these three types of problems (linearly separable, nonseparable, and random noise), we mark these points differently in each plot. Some of the more interesting plots are shown in Figure 1. Notice that all values are plotted on logarithmic scales so points with zero values are outside the plots.

1	no. of feature dimensions
2	no. of points
3	average no. of points per dimension
4	maximum Fisher's discriminant ratio
5	volume of overlap region
6	minimized error by linear programming (LP)
7	% points on boundary (MST method)
8	ratio of average intra/inter class NN distance
9	error of 1NN classifi er
10	nonlinearity of 1NN classifi er
11	error of linear classifi er by LP
12	nonlinearity of linear classifi er by LP
13	no. of iteration in sup. k-means clustering

Table 1. List of investigated measures.

The results indicate that measures 4,5,7,8,9,10,11,12 are useful for describing problem difficulty, since with most of these measures, the linearly separable and the random noise problems do occupy opposite ends of the point distribution. We also observe that there are variable degrees of difficulty among problems of the same type (linearly separable or not, or random noise), e.g., there are linearly nonseparable problems that are almost separable.

In most of the scatter plots with these 8 measures, we observe that the points span a fan-like structure (e.g., plot 7,8). This leads us to believe that at least two factors (possibly more) affect the problem difficulty independently, but their joint effects are most significant. Some measures from the above set are highly correlated (see plots 7,9 and 11,12) as expected due to their similar definitions. These measures, when considered in pairs, provide little additional information. Furthermore, experimental results support the conjecture that (see plot 9,11) linear separability of a problem may not correlate strongly with the nearest neighbor error rate, but it has a distinct effect on the nonlinearity of 1NN classifier (the convex hull effect).

In several of the plots we see the random noise sets appear on the boundary of the fan shape, and they stay far from the real problems due to their exaggerated difficulty (e.g. plot 7,10). This confirms that these real world problems do contain structures that are significantly different from random combinations. Interestingly, in some of these measures the random noise sets span a large range (e.g. 9,10). By checking their appearance in the plots with number of points per dimension, we found that this is due to the apparent simplicity caused by sparseness of samples in the higher dimensional problems. A simple classifier obtained with these apparently easier training sets will turn out to perform very badly on unseen points from the same source.

5. Conclusions

We studied several measures to characterize 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 the selected real world problems is significantly different from that of random noise sets, signifying the existence of learnable structures in such contexts. Also, in this space linear classifiers and nearest neighbor classifiers have very different domains of competence. A challenge is to determine the intrinsic dimensionality of the point distributions in this space, and identify the independent factors.

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 [6], Berlind's hierarchy of indiscernibility [2], Vapnik's VC-dimension theory and his analysis on small sample effects [12], and observations and discussions about several classifiers by Raudys and Jain [8]. An interesting question is the consistency of our chosen measures on bootstrap samples of the training set.

We emphasize that our analysis of the training data can also be applied to their subsets or projections, 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.

Acknowledgements
We thank Eugene Kleinberg and George Nagy for helpful comments.

References

[1] M. Basu, T.K. Ho, The learning behavior of single neuron classifiers on linearly separable or nonseparable input, *Proc. of the 1999 IJCNN*, Washington, DC, July 1999.

[2] R. Berlind, *An Alternative Method of Stochastic Discrimination with Applications to Pattern Recognition*, Doctoral Dissertation, Dept. of Mathematics, SUNY at Buffalo, 1994.

[3] 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.

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

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

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

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

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

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

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

[11] 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.

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

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

[14] 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.

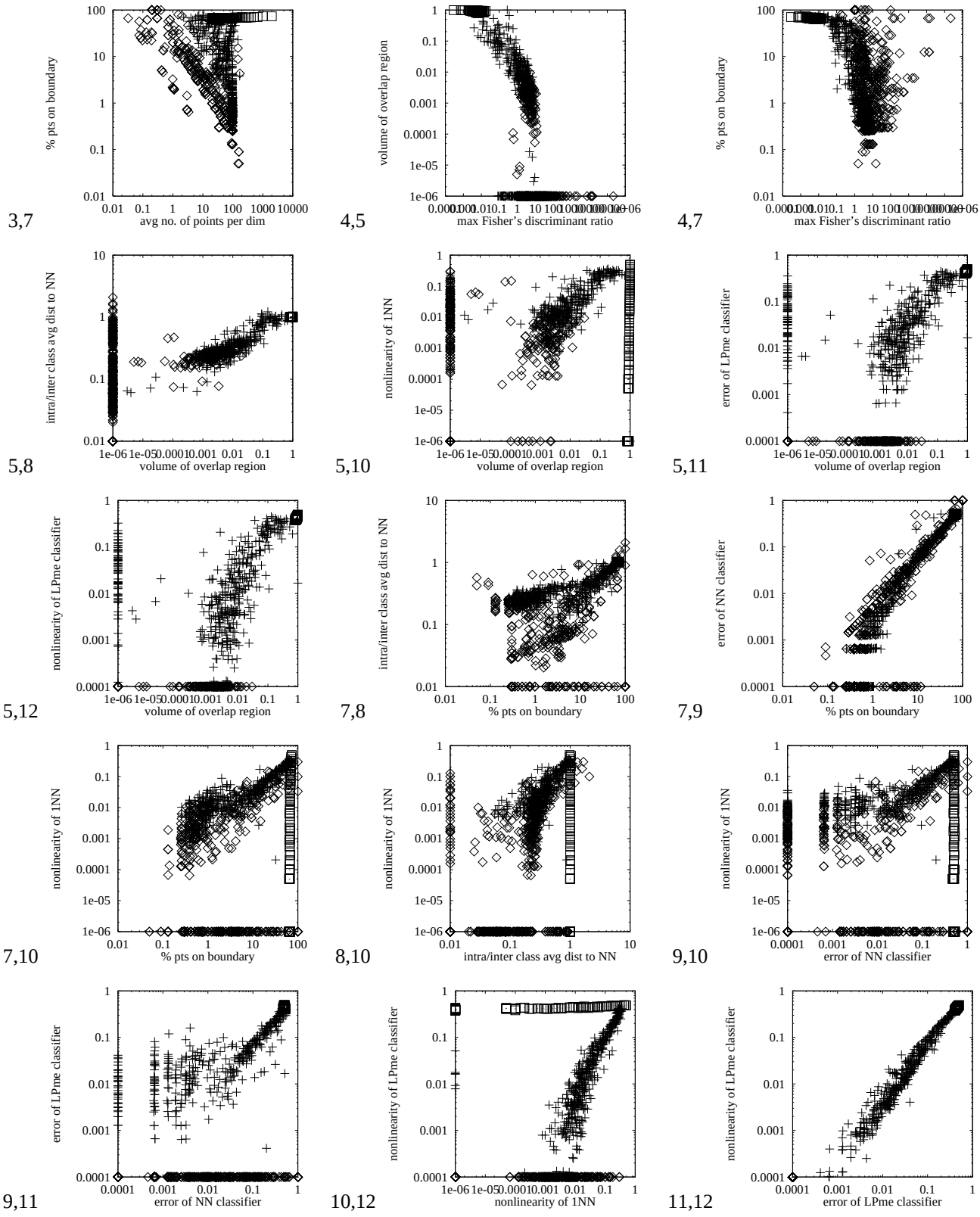

Figure 1. Pairwise plots of selected measures. Markers: diamonds – linearly separable problems; crosses – linearly nonseparable problems; squares – random noise sets.