A Simple Implementation of the Stochastic Discrimination for Pattern Recognition

Dechang Chen[1] and Xiuzhen Cheng[2]

[1] University of Wisconsin – Green Bay, Green Bay, WI 54311, USA
chend@uwgb.edu
[2] University of Minnesota, Minneapolis, MN 55455, USA
cheng@cs.umn.edu

Abstract. The method of stochastic discrimination (SD) introduced by Kleinberg ([6,7])is a new method in pattern recognition. It works by producing weak classifiers and then combining them via the Central Limit Theorem to form a strong classifier. SD is overtraining-resistant, has a high convergence rate, and can work quite well in practice. However, some strict assumptions involved in SD and the difficulties in understanding SD have limited its practical use. In this paper, we present a simple algorithm of SD for two-class pattern recognition. We illustrate the algorithm by applications in classifying the feature vectors from some real and simulated data sets. The experimental results show that SD is fast, effective, and applicable.

1 Introduction

Suppose that certain objects are to be classified as coming from one of two classes, say class 1 and class 2. A fixed number of measurements made on each object form a feature vector q. All the feature vectors constitute a finite feature space $F \subset R^p$. We can classify an object after observing its feature vector q with the aid of the classification rule of SD and a training set $\mathbf{TR} = \{TR_1, TR_2\}$, where TR_i is a given random sample from class i. On an intuitive level, the idea of SD is similar to how people learn: People learn new knowledge and strategies step by step. After years of learning, the knowledge and strategies (or skills) that they have accumulated will enable them to tackle complicated tasks. On a precise mathematical level, the procedure of SD is outlined as follows.

Step 1. Use the training set and rectangular regions to produce t weak classifiers $S^{(1)}$, $S^{(2)}$, ..., $S^{(t)}$, where t is a natural number. See Section 2.

Step 2. For any given feature vector q from F, calculate the average

$$Y(q, \mathbf{S}^t) = \frac{X(q, S^{(1)}) + X(q, S^{(2)}) + \ldots + X(q, S^{(t)})}{t} \quad , \tag{1}$$

where $X(\cdot, \cdot)$ is a base random variable defined later in Section 3.

Step 3. Set a level t classification rule as follows: if $Y(q, \mathbf{S}^t) \geq 1/2$, classify q into class 1; otherwise classify q into class 2. See Section 4.

F.J. Ferri et al. (Eds.): SSPR&SPR 2000, LNCS 1876, pp. 882–887, 2000.

The above procedure follows the idea in [7]. We will study these steps in Sections 2–4.

SD is characterized by the properties of overtraining-resistance, a high convergence rate, and a low misclassification error rate (see [2] and [7]). This will be shown by examples in Section 5. The underlying ideas behind SD were introduced in [6]. Since then, a fair amount of research has been carried out on this method, and on variations of its implementation. See, for example, [1], [2], [3], [4], [7] and [8]. And the results have convincingly shown that stochastic discrimination is a promising area in pattern recognition.

2 How to Produce Weak Classifiers

We produce weak classifiers through resampling. In fact, a weak classifier is s finite union of rectangular regions which satisfies some coverage condition. In this context, a rectangular region in R^p is a region of the form

$$\{\ (x_1, x_2, \ldots, x_p) \mid a_i \leq x_i \leq b_i,\ \text{for } i = 1, 2, \ldots, p\ \}, \tag{2}$$

where a_i and b_i are real numbers for $i = 1, 2, \ldots, p$. Let $\Re$ be an appropriate rectangular region in R^p which contains F. We will utilize those rectangular regions in (2) whose "width" $b_i - a_i$ along the x_i-axis is at least ρ times the corresponding width of $\Re$, where $0 < \rho < 1$ is a fixed constant. The coverage condition is related to a ratio variable r. For any subsets T_1 and T_2 of F, let $r(T_1, T_2)$ denote the ratio of the number of common feature vectors in T_1 and T_2 and the number of feature vectors in T_2. For example, if T_2 contains 5 feature vectors and T_1 and T_2 have 3 feature vectors in common, then $r(T_1, T_2) = 3/5 = 0.6$. It is seen that $r(T_1, T_2)$ represents the coverage of the points in T_2 by T_1.

Now we can define a weak classifier. Roughly speaking, a weak classifier is a finite union of rectangular regions such that the coverage of the points in TR_1 by the union and the coverage of the points in TR_2 by the union are different. Strictly speaking, let β be a fixed real numbers with $0 < \beta < 1$, then an S is said to be a *weak classifier* if S is a union of at most κ rectangular regions in R^p which satisfies $|r(S, TR_1) - r(S, TR_2)| \geq \beta$. The condition $r(S, TR_1) \neq r(STR_2)$ simply states that S can actually be used as a (very weak) classifier. To illustrate this, consider an S, which contains 40 sample points from TR_1 and 60 from TR_2. Then $r(S, TR_1) = 40/n_1$ and $r(S, TR_2) = 60/n_2$. Suppose $40/n_1 > 60/n_2$. That is, the coverage of the points in TR_1 by S is greater than the coverage of the points in TR_2 by S. Let S^c denote the complement of S in F. Then since $1 - 40/n_1 < 1 - 60/n_2$, one sees that the coverage of the points in TR_2 by S^c is greater than the coverage of the points in TR_1 by S^c. Thus intuitively we could use S to classify F by deciding that any sample point q from S belongs to class 1 and any other sample point q from S^c belongs to class 2. This of course gives a (very) weak classifier.

3 Base Random Variables

To connect weak classifiers with feature vectors, we need a *base random variable* $X(\cdot,\cdot)$. Given a feature vector q and a weak classifier S, the value of X is defined to be

$$X(q,S) = \frac{1_S(q) - r(S,TR_2)}{r(S,TR_1) - r(S,TR_2)} \quad , \tag{3}$$

where $1_S(q) = 1$ if q is contained in S and 0 otherwise. $X(\cdot,\cdot)$ in (3) can be understood as the standardized version of $1_S(q)$, which gives the simplest way to connect the weak classifiers S and feature vectors q.

4 Classification Rule

Let $\mathbf{S}^t = (S^{(1)}, S^{(2)}, \ldots, S^{(t)})$ be a random sample of t weak classifiers. For any $q \in F$, define $Y(q,\mathbf{S}^t)$ as stated in (1). Under some mild conditions, the Central Limit Theorem can be used to show the following fact. If t is large enough then there is a high probability that $Y(q,\mathbf{S}^t)$ is close to 1 for any q from TR_1 and close to 0 for any q from TR_2 (see Theorem 1 in [2].). Hence one can define the following

Level t Stochastic Discriminant Classification Rule: For any $q \in F$, if $Y(q,\mathbf{S}^t) \geq 1/2$, classify q into class 1, otherwise classify q into class 2.

5 Experimental Analysis

In this section, we report the experimental results on classifying feature vectors from several problems. The emphasis will be placed on normal populations. The comparison of SD with non-SD methods is also given.

Example 1 (Two normal populations with equal covariance matrix). Consider two distributions $N(\boldsymbol{\mu}_1,\ \boldsymbol{I})$ (class 1) and $N(\boldsymbol{\mu}_2,\ \boldsymbol{I})$ (class 2), where $\boldsymbol{I}$ is the 2×2 identity matrix, $\boldsymbol{\mu}_1$ is the vector $(1.5, 0)'$, and $\boldsymbol{\mu}_2$ is the vector $(0, 0)'$. Both of the prior class probabilities π_1 (for class 1) and π_2 (for class 2) are equal to 1/2. The training set contains 400 points from each class, and test set contains 1000 points from each class. Let $\Re_1$ be the smallest rectangular region which contains both training and test data. Suppose $\lambda \geq 1$. Let $\Re_\lambda$ denote the rectangular region similar to $\Re_1$ whose center is the same as that of $\Re_1$ and whose "width" along the x_i-axis is λ times the corresponding width of $\Re_1$. We regard $\Re_\lambda$ as our $\Re$ defined in Sect. 2. For the resampling process, $\lambda = 1$, $\rho = 0.3$, $\beta = 0.52$, and $\kappa = 5$. The test error from SD is below 23.55% when the level $t \geq 4700$. See Fig. 1 for the performance of SD. From the figure, we see that both training and test errors start to decrease at the beginning and then quickly level off, forming two "parallel curves", as more weak classifiers are added. This phenomenon is common for SD classification procedures. As a comparison, the linear discriminant rule yields a test error 23.15%.

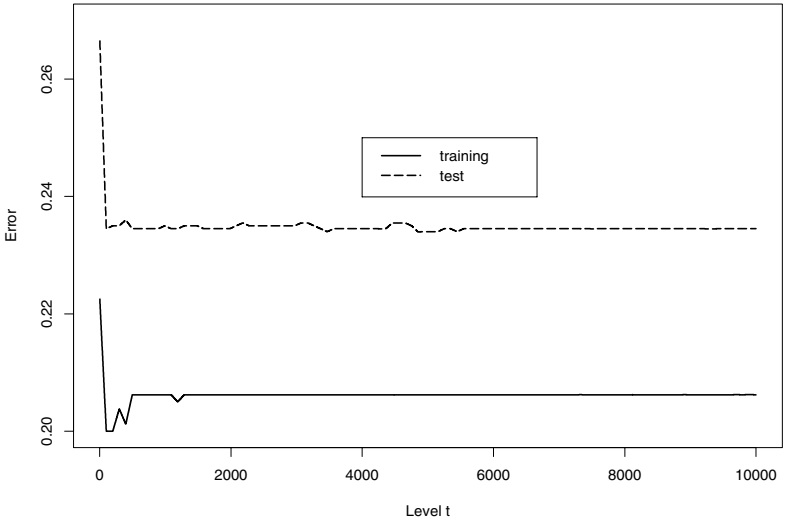

Fig. 1. Training and test errors for the classification with two normal distributions having the same covariance matrix. The training set contains 400 points from each class, and the test set contains 1000 points from each class.

Example 2 (Classifying Alaskan and Canadian salmon). Here we will show the performance of SD on the salmon dataset from Johnson and Wichern ([5]). The original data contain the information of gender, diameter of rings for the first-year freshwater growth, and diameter of rings for the first-year marine growth for 50 Alaskan salmon and 50 Canadian salmon. We treat both freshwater and marine growth ring diameters as the features of salmon.

Johnson and Wichern ([5]) note that the data appears to satisfy the assumption of bivariate normal distributions, but the covariance matrices may differ. Thus the usual quadratic classification rule may be used to classify the salmon. Using the quadratic rule and equal prior probabilities, the error rate from a 10-fold cross-validation is then 8%.

To apply SD, we set $(\lambda, \rho, \beta, \kappa) = (1.005, 0.1, 0.5, 5)$. From the same data sets as those used with the quadratic rule, the 10-fold CV error rate from SD is virtually 9%. See Fig. 2 for the details.

Other comparisons of SD with non-SD methods are also available. For example, [2] considers the classification for the Pima Indians diabetes dataset described in [9]. The test error from SD is actually identical to the best result in [9]. Section 3 of [7] reports one experiment on handwritten digit recognition and another experiment on classifying boundary types in DNA sequences. In the first experiment, SD is compared with a nearest neighbor algorithm, a k-nearest neighbor algorithm, and a neural network. In the second experiment, SD is com-

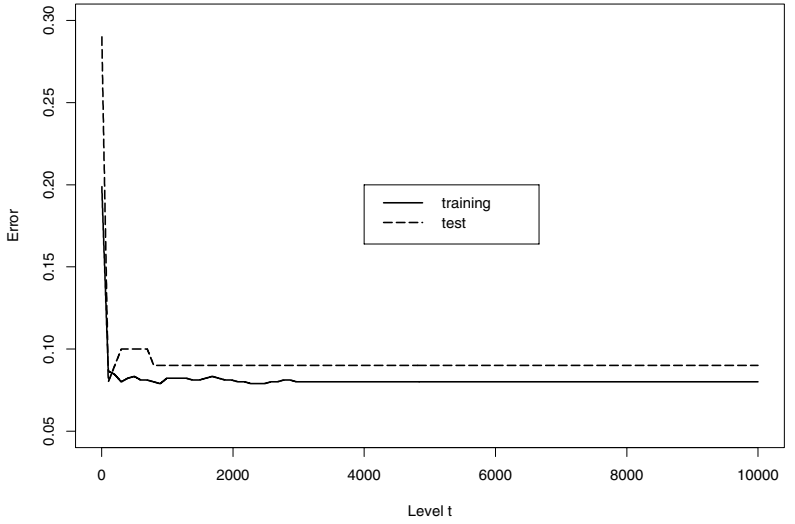

Fig. 2. The averaged training and test errors for classifying Alaskan and Canadian salmon. The errors are based on a 10-fold cross-validation procedure.

pared with more than 20 different methods. In both cases comparisons show that SD yields the best test set performance.

Notes. When applying SD to a dataset, we need the values of λ, ρ, κ, and β. From Sect. 2, we know that these 4 parameters together determine weak classifiers. Since the quantitative relationship among these parameters is not available, we can apply SD to the training set alone to find out the combination of these parameters with which the misclassification rate for **TR** is minimum. Thus, we propose the following two–step procedure. First, we run SD over **TR** by stepping through the range of these parameters and find out the combinations corresponding to the best achieved **TR** performance. Since SD has an exponential convergence rate ([2]), this step is practical. In fact, usually we can obtain several satisfactory combinations and we choose the one with which SD runs fastest. Then, we apply SD to both training and test sets with the selected parameters.

Acknowledgments. The first author is very grateful to his advisor Professor Eugene Kleinberg for providing an introduction to the field of pattern recognition and sharing with the author his own research.

References

1. Berlind, R.: An Alternative Method of Stochastic Discrimination with Applications to Pattern Recognition. Doctoral Dissertation, Dept. of Mathematics, State University of New York at Buffalo (1994)
2. Chen, D., Huang, P., Cheng, X.: A Concrete Statistical Realization of Kleinberg's Stochastic Discrimination for Pattern Recognition, Part I. Two - Class Classification. Submitted for Publication
3. Ho, T. K.: Random Decision Forests. In: Kavanaugh, M., Storms, P. (eds.): Proceedings of the Third International Conference on Document Analysis and Recognition. IEEE Computer Society Press, New York (1995) 278-282
4. Ho, T. K.: The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence **20** (1998) 832-844
5. Johnson, R. A., Wichern, D. W.: Applied Multivariate Statistical Analysis. 4th edn. Prentice Hall (1998)
6. Kleinberg, E. M.: Stochastic Discrimination. Annals of Mathematics and Artificial Intelligence **1** (1990) 207-239
7. Kleinberg, E. M.: An Overtraining-Resistant Stochastic Modeling Method for Pattern Recognition. Annals of Statistics **24** (1996) 2319-2349
8. Kleinberg, E. M., Ho, T. K.: Building Projectable Classifiers of Arbitrary Complexity. In: Kavanaugh, M. E., Werner, B. (eds.): Proceedings of the 13th International Conference on Pattern Recognition. IEEE Computer Society Press, New York (1996) 880-885
9. Ripley, B. D.: Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge (1996)