NASA-CR-204201

653

NDB
7N-65-CR
027674

PARTIALLY SUPERVISED CLASSIFICATION WITH OPTIMAL SIGNIFICANCE TESTING[1] *

Byeungwoo Jeon and David A. Landgrebe

School of Electrical Engineering
Purdue University, W. Lafayette, IN 47907, U.S.A.
Telephone: (317) 494-3486, FAX : (317) 494-6440

ABSTRACT

This paper addresses the problem of estimating an optimal acceptance probability to be used for significance testing as applied to partially supervised classification where the class definition and corresponding training samples are provided *a priori* only for one specific class of interest. Considering the effort in both time and man-power required for a well-defined, exhaustive list of classes with their representative training samples even if there is just one class of interest to identify, the "partially" supervised capability would be very desirable, assuming adequate classifier performance can be obtained. The optimal acceptance probability is estimated directly from the data set. Experiments with both simulated and real data show very satisfactory results.

Key Words: partially supervised classification, significance testing, acceptance probability, object detection

I. INTRODUCTION

In practical applications of pattern classification techniques, it is not unusual to confront "partially" supervised problems [1,2] where one is to identify only a particular class of objects among others with only statistical information pertaining to those objects of interest. It can arise in such cases where defining all the classes and collecting their statistical information is impossible or very expensive in terms of time and manpower. Design of a conventional supervised classifier requires training samples for all the classes in the given data in order to insure optimal performance. Considering the effort in both time and man-power required to have a well-defined, exhaustive list of classes with a corresponding representative set of training samples, the partially supervised capability would be very desirable, assuming adequate classifier performance can be obtained. We addresses the partially supervised classification problem especially when the class definition and its statistical characterization are provided *a priori* for only just one particular class.

Significance test is useful in such problems as thresholding, or the single hypothesis testing [3,4] which is very similar to partially supervised classification. One important element in significance testing is the acceptance probability (or significance level), which is usually set by the data analyst in such a way that the type I (*i.e.*, omission) error rate is kept within a pre-specified level. The commission error, or type II error cannot be easily evaluated unless the relative distributions of all classes in the data set are available. Note, however, that a mixture density estimate can give an estimate of the probability density of "others" class. This paper presents an algorithm which can automatically estimate the optimal acceptance probability from the given data set, without the user's supervision, under selected optimality conditions such as the class-averaged classification error, or the generalized total classification error, to be used with significance testing for the partially supervised classification.

[1]This work was funded in part by NASA under grant NAGW-925

II. SIGNIFICANCE TESTING

Suppose there is a data set, $X \equiv \{x_1, \cdots, x_N\}$ with N samples. x_i's are q-dimensional feature vectors. One is only interested in identifying a single class, denoted by C_{int}, *i.e.*, discriminating between it and the "others" class, denoted by C_{others}. Let $f_x(x|C_{int})$ be the probability density function (PDF) of the class of interest. It is assumed either to know the PDF of the class of interest, or, to have a representative training samples from which a density estimate can be made. The prior probabilities of the class of interest and others are indicated by π_{int} and π_{others}. Even though the following derivations do not require any specific family of PDF for the class of interest, multivariate normality will be assumed for simplicity's sake. It is straightforward to generalize to other PDF's. Furthermore, without loss of generality, the class of interest is assumed to have zero mean, and an identity covariance matrix.

With $f_x(x|C_{int})$ being the standard multivariate Gaussian, a natural choice for the test statistic would be $T(x) = x^T x$. The availability of the statistical characterization of the class of interest enables control of the omission error, denoted by $\varepsilon_1(\alpha)$. The value, $(1-\alpha)$ defines the maximum allowable omission error and is often called the *significance level*. The parameter α will be called the *acceptance probability*. The commission error, denoted by $\varepsilon_2(\alpha)$, is generally difficult to control since its evaluation requires unavailable statistical knowledge about all alternatives. To avoid potentially excessive omission or commission errors, the acceptance probability must be carefully determined by checking the relative distribution of data samples with respect to the class of interest. An automatic estimation capability of optimum acceptance probability is thus very desirable.

In significance testing which can be viewed as a problem of *single* hypothesis testing, the optimal acceptance probability can be obtained in a similar way to the simple binary hypothesis counterpart [5] for which an optimal Bayes minimum expected cost test is well known as the likelihood ratio test whose design requires functions of omission and commission errors. The expression for commission error in significance testing, is ordinarily not readily available *a priori*. Nevertheless, estimating the commission error function for a given data set is possible, as will be discussed in following section from the mixture density estimate. With the estimated commission error function, the same idea of simple binary hypothesis testing mentioned above can be applied also to significance testing in estimating the optimal acceptance probability.

III. OPTIMAL SIGNIFICANCE TESTING

Suppose there are N_1 samples belonging to C_{int} in the set X. N_1 is generally unknown. $N(\alpha)$, the *expected* number of samples in X accepted with the acceptance probability α, is,

$$N(\alpha) \equiv N \int_0^{\lambda_\alpha} f_Y(s)\, ds, \quad 0 \le \alpha \le 1$$

*Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS'93), Tokyo, pp 1370-2, August 1993.

where $f_Y(y)$ is the mixture PDF of y, $y = x^T x$, and λ_α is the threshold corresponding to acceptance probability α. In the most simplistic way, $N(\alpha)$ can be obtained by counting the number of samples accepted while varying α. Even though it is quite simple and fast enough, it may have some drawbacks coming from the counting nature. To avoid the drawbacks, a kernel-based Parzen density estimate [6] of $f_Y(y)$ can be obtained first to compute $N(\alpha)$. Note that y's are all non-negative and the domain of the density estimation is $[0,+\infty)$. In this case, the use of a symmetric kernel function such as the Gaussian kernel would result in underestimation near zero since there are no samples with negative values. This underestimation can be avoided by using positive reflection [7] in which a new density estimate is obtained with an augmented set of y's.

$N_1(\alpha)$, defined as the expected numbers of data points of C_{int} accepted with α, is given as $N_1(\alpha) = \alpha N_1$ and the omission error rate is $\varepsilon_1(\alpha) = 1 - \alpha$. The commission error rate is given as,

$$\varepsilon_2(\alpha) = \frac{N(\alpha) - \alpha N_1}{N - N_1}$$

Note that $\varepsilon_1(\alpha)$ is strictly decreasing with slope -1. $\varepsilon_2(\alpha)$ is a monotonically increasing function with increasing rate dependent on $N(\alpha)$. Evaluating $\varepsilon_2(\alpha)$ generally requires knowledge of N_1.

The optimal test can be designed by minimizing the a posterior expected cost given as [5],

$$E(\alpha) = A\, \pi_{int}\, \varepsilon_1(\alpha) + \pi_{others}\, \varepsilon_2(\alpha) \qquad (1)$$

No cost is assumed for a correct decision and the constant A, $A > 0$, is the cost on making *omission* error relative to the cost of making *commission* error being 1. The optimal acceptance probability is dependent on the optimality criterion. Note that different error criteria can be represented with eq. (1) by substituting different A values. The class-averaged classification error criterion which is a simple average of omission and commission errors can be represented equivalently by eq. (1) with $A = \pi_{others}/\pi_{int}$. The Bayesian total probability of error criterion (with A=1) minimizes the weighted sum of omission and commission errors with the prior probabilities. Equation (1) will be called the "generalized" total classification error criterion. The optimal acceptance probability can be obtained by checking the elements in the set S defined as,

$$S \equiv \{\, \alpha \mid \frac{dE(\alpha)}{d\alpha} = 0 \ \text{ and } \ \frac{d^2E(\alpha)}{d\alpha^2} > 0,\ 0 \le \alpha \le 1 \}$$

Each element in S will correspond to the (local) minima of eq. (1). The global minimum can be selected by comparing the actual values of $E(\alpha)$ at different α's in S in the following way. Suppose α_i, α_j are elements in S, then, the difference in error can be written as,

$$E(\alpha_i) - E(\alpha_j) = \frac{\Delta_{ij}}{N} \qquad (2)$$

$$\text{where, } \Delta_{ij} \equiv \left[N(\alpha_i) - N(\alpha_j) - (\alpha_i - \alpha_j) \cdot (1+A) \cdot N_1 \right]$$

By checking the signs of the Δ_{ij}'s, the acceptance probability which attains global minimum can be selected from the set S. Notice that evaluating eq. (2) requires π_{int}, but, in the case of the class-averaged classification error criterion, it can be evaluated even without the prior probability since substituting $A = \pi_{others}/\pi_{int}$ results in a quantity independent of π_{int}. This property of the class-averaged classification error criterion will be very useful in actual application of this algorithm since the prior probabilities are unknown in most problems.

IV. EXPERIMENTS AND DISCUSSION

To test the performance of the proposed estimating algorithm for optimal acceptance probability in significance testing and application to the partially supervised classification, experiments were carried out with both simulated and real data.

A. Experiment with Simulated Data

For a test with simulated data, 1000 samples were generated for the class of interest to be bivariate Gaussian with zero mean and an identity covariance matrix. For the class "others", 2000 samples were generated to be bivariate Gaussian with a mean $[d,0]^T$, $d > 0$, and an identity covariance matrix. With this set-up, the exact amount of overlap between the two distributions can be calculated. The term "overlap" is defined as the volume which is shared by the two probability density functions [1]. By varying d, data sets with different degrees of separability can be simulated. d was increased from 0.1 to 5 in steps of 0.1. If d = 0.1, there is 96.02% of overlap between the two distributions, and in the case of d=5, there is only 1.24% of overlap. To avoid any random error due to the data generation process and its effect on evaluating the experimental result, data sets were generated 50 times with different seed numbers, and the averaged result was used in comparison.

To make a comparison with the *estimated* values, optimal acceptance probabilities were *manually* selected by changing acceptance probability from 0.01 to 0.99 in steps of 0.01 and choosing the best one based on the selected optimality criterion. These manually selected were denoted by "scanned" values and compared with the estimates obtained by the proposed algorithm. The estimated acceptance probabilities with both the class-averaged and the total classification error criteria are shown in Fig. 1. When applying the total classification error criterion, the true value of prior probabilities were used.

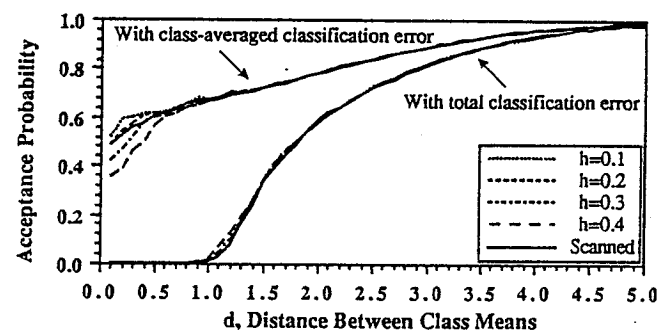
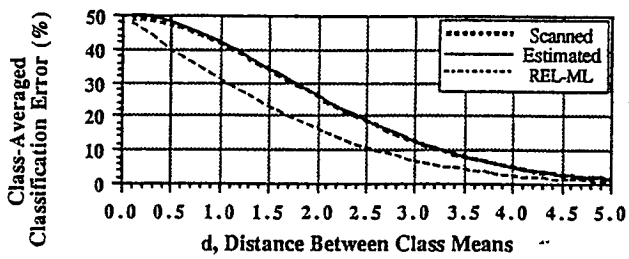


Figure 1. Estimated optimal acceptance probability versus d, the distance between two class means; Solid lines show the manually selected acceptance probabilities. Dotted lines show the estimated optimal acceptance probabilities using the proposed method. h is the Parzen window size.

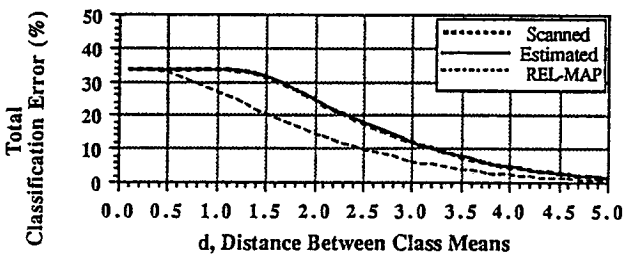
The density estimate required for $N(\alpha)$ was obtained by employing a Gaussian Kernel-based Parzen density estimate with data set augmented by positive reflection [7]. In Fig. 1, the estimated values followed very closely those manually selected especially when d was large. The optimal acceptance probability based on the total classification error was observed to be near 0 when d was not large enough, since the total classification error was an increasing function of acceptance probability for those small d values. When $d < 1.0$ under the class-averaged classification error criterion, some degree of difference was observed between the estimated and the manually selected. Since the curve of class-averaged classification error was nearly flat when $d < 1.0$ [1], an exact location of the minimum of the class-averaged classification error was hard to pinpoint and thus, there was observed a relatively large standard deviation not only in the estimated but also in the manually selected optimum values [1]. In spite of those discrepancies in estimated optimal values, there

was not much difference in the resulting class-averaged classification errors. Since less than 1% of the differences were observed under both optimality criteria by varying the window size, classification results are shown only for h=0.2 in Fig. 2.

The significance testing deals with only the values of the test statistic, therefore there is a dimensionality reduction of feature vectors to one-dimensional space of the selected test statistic, and this causes information loss in classification.



(a)



(b)

Figure 2. Classification error versus d, the distance between two class means; (a). with class-averaged classification error criterion (b). with total classification error criterion; "REL-ML" and "REL-MAP" are results with the relative maximum likelihood and maximum a posterior classifier respectively. Parzen window size h was 0.2.

To see the effect of dimensionality reduction, a (relative) maximum likelihood classifier (denoted as "REL-ML") and a maximum a posterior classifier (denoted as "REL-MAP") were designed in the original q-dimensional space with known class statistics of C_{int} and C_{others}. Their classification results were also included in Fig. 2. Under both optimality conditions, the estimated optimal acceptance probabilities resulted in almost the same performances with manually selected values. There was a maximum of about 12% error increase due to the dimensionality reduction.

B. Experiment with Real Data

For a test with real data, a Landsat Thematic Mapper data set which was acquired over an agricultural area in Tippecanoe County, Indiana in July was used with all seven features (i.e., q=7). From the available ground truth data, four different information classes - corn, soybeans, wheat and alfalfa/oats - were identified. About 10% of the samples were randomly selected from each information class to be served as training samples. Based on clustering, two sub-classes were found for corn and soybeans and four sub-classes were found for wheat and alfalfa/oats.

The class-averaged classification error criterion was applied to the sub-classes of the selected information class. A Gaussian kernel function was used in density estimation with the positive reflection. Various window sizes (h=0.1 ~ h=0.6 in steps of 0.1) were tried to observe virtually no differences. The results reported here were obtained with h=0.5. In order to make a comparison with the *estimated* optimal acceptance probability, a specific value which attained the given optimality condition was *manually* determined as before. It is denoted by "*scanned*," in Fig. 3. The estimated optimal acceptance probabilities using the proposed method agreed quite well with those manually selected [1]. The class-averaged classification errors with estimated optimum acceptance probability were also very close to those obtained with the manually selected values. The maximum difference between the estimated and the manually selected was only 0.03, except for the sub-class 4 of "alfalfa/oats" which had a difference of 0.08. The corresponding difference in the class-averaged classification error in this sub-class was only 0.34%.

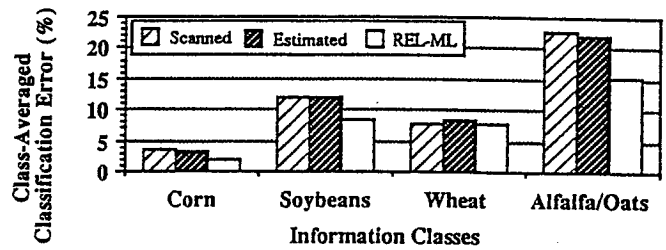


Figure 3. Comparisons of class-averaged classification error with optimal acceptance probabilities estimated with the class-averaged classification error criterion. "REL-ML" is the result obtained with a (relative) maximum likelihood classifier designed with all 12 sub-classes.

The class-averaged classification errors evaluated for each information class are compared in Fig. 3. With all 12 sub-classes and their class statistics, a relative maximum likelihood classifier in the original seven dimensional space was designed and its result (denoted by "REL-ML") is also included in Fig. 3 to show the effect of dimensionality reduction. In the corn and wheat classes, there seemed to be not much information loss due to dimensionality reduction. However, there was as much as 3 ~ 5% of class-averaged classification error increase in soybeans and alfalfa/oats.

V. CONCLUSIONS

In this paper, a problem of estimating optimal acceptance probability in significance testing was addressed to be applied for partially supervised classification. As the optimality criteria, both class-averaged classification error and generalized total classification error criteria were considered. If the class of interest doesn't need to be decomposed into sub-classes, the class-averaged classification error criterion can be applied even without the knowledge of prior probabilities. The automatic estimation procedure will make the popular significance testing technique even more powerful in many fields of its application including partially supervised classification problem.

REFERENCES

[1] B. Jeon and D. A. Landgrebe, "Design of Partially Supervised Classifiers for Multispectral Image Data," Technical Report, TR-EE 93-11, School of Electrical Engineering, Purdue University, 1993

[2] B. Jeon and D. A. Landgrebe, "Absolute Classification with Unsupervised Clustering," Proc. of IEEE Intern. Geos. and Remote Sensing Symp., pp.1609-1611, 1992

[3] C. W. Therrien, *et al.*, "Statistical Model-Based Algorithms for Image Analysis," Proc. of IEEE, Vol.74, No.4, pp.532-551, April, 1986

[4] K. Fukunaga, *et al.*, "The Acquisition Probability for a Minimum Distance One-class Classifier," Trans. IEEE Aero. and Electronic Systems, AES-23, pp.493-499, 1987

[5] H. L. Van Trees, Detection, Estimation, and Modulation Theory, Part I, John Wiley & Sons, New York, 1968

[6] B. W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman and Hall, 1986

[7] L. I. Boneva, *et al.*, "Spline transformations: three new diagnostic aids for the statistical data-analysis (with Discussions)," Journal Royal Statist. Soc. B, 33, pp.1-70, 1971