PARTIALLY SUPERVISED CLASSIFICATION USING WEIGHTED UNSUPERVISED CLUSTERING

Byeungwoo Jeon[1] and David A. Landgrebe[2]

[1] School of Electrical and Computer Engineering
Sung Kyun Kwan University, Korea

[2]School of Electrical and Computer Engineering
Purdue University, W. Lafayette, IN 47907-1285, U.S.A.

ABSTRACT

This paper addresses a classification problem in which class definition through training samples or otherwise is provided a priori only for a particular class of interest. Considerable time and effort may be required to label samples necessary for defining all the classes existent in a given data set by collecting ground truth or by other means. Thus, this problem is very important in practice, because one is often interested in identifying samples belonging to only one or a small number of classes. The problem is considered as an unsupervised clustering problem with initially one known cluster. The definition and statistics of the other classes are automatically developed through a weighted unsupervised clustering procedure that keeps the known cluster from losing its identity as the "class of interest." Once all the classes are developed, a conventional supervised classifier such as the maximum likelihood classifier is used in the classification. Experimental results with both simulated and real data verify the effectiveness of the proposed method.

Key Words: partially supervised classifier, one-class classifier, single hypothesis testing, unsupervised clustering, significance testing

I. INTRODUCTION

In the real world, there are many applications where a classifier that can recognize only a single class of samples is sufficient. Since the process of gathering training samples or otherwise labeling training samples is very expensive in terms of time and manpower, it would be very useful if one can design such a classifier given only the training samples belonging to the class of interest.

We call this a "partially supervised classification" in the sense that the prior information is available only for the class of interest. It is assumed that one either knows the probability density function of the class of interest, or has training samples from which its unknown density function can be estimated. It is also known as the single hypothesis problem [1,2]. This kind of problem arises in such cases where defining all the classes and gathering corresponding statistical information is impossible or very expensive in terms of time and manpower. Examples of possible applications include target detection [1], object detection out of various backgrounds [3], texture detection, and cloud identification.

The maximum likelihood (ML) classifier, one of the most widely used relative classifiers, is not a good choice here since the relative comparison of log-likelihood values requires training samples for all other classes in order to adequately train the classifier. The necessity of training samples for, or otherwise defining all other classes (i.e., fully supervised) is an onerous shortcoming especially when there are large numbers of classes and/or features to deal with. Note that the necessary number of training samples is dependent on the number of features and the number of classes [4], and insufficient training samples compared to the number of features can degrade the classification performance [5].

On the other hand, classifiers such as the parallelepiped classifier or a scheme based upon a known absorption feature for a specific material, classify data samples on an absolute basis without regard to the spectral responses of other materials or classes that may be in the scene. Therefore design of such a classifier requires class definition through training samples only for the particular class under consideration. When properly designed, a relative classifier such as the ML classifier, nearly always provides better performance, and is much less sensitive to many unmanageable factors, e.g., atmospheric conditions, calibration, etc. However, the operational simplicity of the absolute scheme (such as the parallelepiped classifier) may make it the scheme of choice in many practical instances.

In this paper, we seek both advantages of the reduced requirements for necessary prior knowledge in the absolute scheme and the potentially robust and powerful discriminating capability of the relative scheme by developing an automatic mechanism for extracting statistical information corresponding to the others class without recourse to prior knowledge supplied by the data-analyst. We formulate this as a special case of unsupervised clustering with one particular cluster initially known. Its key problem is how to define and find statistics for the other clusters without confusion between the class of interest and the others class.

The proposed method is in three steps. First, each data sample is assigned a weight factor indicating likelihood of being from "the others" class. The second step is to develop the initial definition of clusters corresponding to the others class using the weighted unsupervised clustering. Finally the cluster statistics are iteratively refined, and a conventional relative classifier such as the maximum likelihood classifier makes decisions using the cluster statistics.

The organization of this paper is as follows. Section II presents a formal statement of the partially supervised classification problem with a brief review of previous works. Section III discusses the weighted unsupervised clustering procedure for the unknown initial class definitions and the subsequent class statistics development through clustering. Experimental results with both simulated and real LANDSAT Thematic Mapper (TM) data are presented with discussion in Section IV. Finally, some observations and concluding remarks in Section V complete this paper.

II. PARTIALLY SUPERVISED CLASSIFICATION

Suppose samples belonging to a particular class of interest need to be identified from a given data set, $\mathbf{X} \equiv \{x_1, \text{---}, x_N\}$. Each data sample has a q-dimensional feature vector x_i. The number of samples in the whole data set $\mathbf{X}$ is N. N_1 denotes the number of samples in $\mathbf{X}$ which belong to the class of interest samples; it is *unknown*. Let us denote the class of interest and the class of "the others" by C_{int} and C_{others}, respectively. C_{others} might consist of several sub-classes none of which are of interest. C_{int} is an information class [4] which correspond to a physically meaningful entity. In the derivation of the proposed method, C_{int} is assumed to be modeled by a known probability density function(PDF) denoted by $f_x(x|C_{int})$. If the PDF is not known a priori, one can choose a proper family of PDF's and estimate its parameters using the given training samples belonging to the class of interest.

In some cases, more than one PDF may be necessary to model the distribution of the class of interest. For example, suppose the selected PDF is (uni-modal) Gaussian, but the distribution is multi-modal. In such cases, one can sub-group the data set $\mathbf{X}$ first and apply the proposed method to each sub-group separately; suppose K PDF's are required for the class of interest. Then, one divides the whole data set $\mathbf{X}$ into K sub-groups by classifying (e.g., using ML Gaussian classification) the samples in $\mathbf{X}$ into the K sub-classes each of which is characterized by one of the K PDF's. In each of the K sub-groups, note that the class of interest is modeled by one PDF. Furthermore, $f_x(x|C_{int})$ is assumed to have zero mean and an identity covariance matrix. This causes no loss of generality, since, if not, it is always possible to normalize the data x to be so by a straightforward linear operation of $\Sigma^{-\frac{1}{2}}(x - M)$ [2] where Σ and M are respectively the covariance and mean of x.

One straightforward way of partially supervised classification is by the significance testing [6] under which the null hypothesis, H_0: x $\quad C_{int}$ is tested against all other alternatives with a test: Reject H_0 if $T(x|C_{int}) \equiv f_x(x|C_{int}) < \lambda_\alpha$. It amounts to a simple thresholding where the threshold λ_α is determined in such a way that the maximum rejection probability (i.e., omission error) is not more than the significance level α. The significance level should be decided based on the probability distribution overlap between the class of interest and the others. Since the user does not have such prior knowledge to determine an optimal significance level, the success of this approach is very limited. Furthermore, the reduction to one dimensional space of the test statistic causes the loss of much information valuable in classification [1,2].

The approach in [7] avoids the difficulties by iteratively estimating the prior probability of C_{int} and the PDF of C_{others} from the Parzen estimate of the mixture PDF through the EM (Expectation and Maximization) algorithm [8]. The mixture PDF $f_X(x_i)$ is written as $f_X(x_i) = \pi f_X(x_i \mid C_{int}) + (1-\pi)\, f_X(x_i \mid C_{others})$, $x_i \quad \mathbf{X}$. Although this can avoid the information loss due to dimensionality reduction and the user's guess of an appropriate significance level, it is computationally very intensive.

In this paper, we approach the problem in the context of unsupervised clustering [2]. This approach differs from general unsupervised clustering in that (1) one is interested in finding samples of only one particular cluster (or class) and one has its statistical information such as the probability density function a priori; (2) the clusters corresponding to the others class do not need to be meaningful as useful informational classes and, furthermore the confusion between those clusters are not important as long as they are differentiable from the class of interest.

The mixture density $f_X(x)$ is written as a weighted sum of *L* probability density functions as,

$$f_X(x) = \sum_{k=1}^{L} \pi_k \quad f_X(x \mid C_k) \tag{1}$$

where π_k and $f_X(x|C_k)$ are respectively the prior probability and probability density function of the k$^{\underline{th}}$ class, *k* = 1, --- , *L*, and $\pi_1 + \text{---} + \pi_L = 1$. The notation of C_1 and C_2, ---, C_L means that $C_1 = C_{int}$ and C_2, ---, C_L are the sub-classes of C_{others}. According to the assumption, only $f_X(x|C_1)$ is known a priori. We develop the probability distribution functions corresponding to the others class through unsupervised clustering so that the time-consuming density estimation process in [7] can be avoided. In the process, it is important to ensure that there is no significant confusion between C_{int} and the clusters corresponding to C_{others} so that the cluster statistics of C_{others} should not be biased by the samples belonging to C_1. One conceivable approach for reducing the bias is to find the clusters of C_{others} by performing clustering with a subset of data in which a significant portion of the C_1 samples are removed through the significance testing. In addition to the difficulty in selecting the proper significance level, the approach still suffers the biasedness problem especially when C_{others} is not well separated from $C_{1.}$ Instead of removing the effect of C_1 samples in a rather absolute way, we assign to each sample a weight factor which indicates the relative likelihood of belonging to C_{others} and determine the number of clusters *L* and the unknown cluster statistics using a new unsupervised clustering with the weights.

Once the initial specifications of the clusters are obtained through unsupervised clustering with weights, then, a conventional supervised clustering procedure iteratively refines the unknown class statistics. The class statistics developed are used in the relative classification scheme chosen. The proposed procedure is summarized in Fig.1.

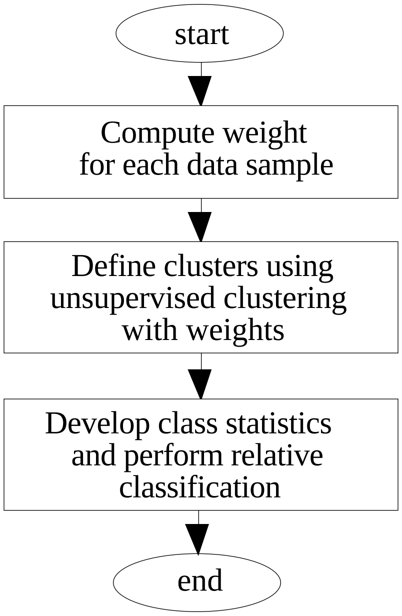
start
Compute weight for each data sample
Define clusters using unsupervised clustering with weights
Develop class statistics and perform relative classification
end

Figure 1. Flowchart of proposed partially supervised classification method

III. PARTIALLY SUPERVISED CLASSIFICATION USING UNSUPERVISED CLUSTERING WITH WEIGHTS

A. Computation of Weights

Initial specifications of clusters which can initiate the supervised clustering are found through an unsupervised clustering procedure. In creating a new cluster, setting up a proper condition of new cluster creation is important. If too many small clusters are generated near the origin in the feature space (that is, the class of interest), these will take up significant portion of samples from C_{int}. To reduce the sensitivity of the initial cluster specification on the cluster creation parameter, each data point x_i is assigned with a weight $\overline{w}_{i1}$ in eq.(2.a) which is the relative likelihood of not belonging to C_{int}.

$$\overline{w}_{i1} \equiv 1 - w_{i1} \tag{2.a}$$

$$\text{where, } w_{i1} = \pi_1 \frac{f_X(x_i \mid C_1)}{f_X(x_i)} = \frac{N_1 \; f_X(x_i \mid C_1)}{N \; f_X(x_i)} \tag{2.b}$$

Note that evaluating the weight factor, $\overline{w}_{i1}$, requires π_1 (or N_1 since $\pi_1 = N_1 / N$, where N_1 is the number of samples in **X** belonging to the class of interest) and the mixture density $f_X(x_i)$. Since the purpose of the unsupervised clustering is to provide initial specification of clusters to launch the clustering process and a direct estimation of $f_x(x_i)$ through non-parametric density estimation would require complex computation, practical approximation is made by noting that w_{i1} can be expressed as a ratio,

$$w_{i1} = \frac{N_1 \, f_x(x_i \mid C_1)\Delta V}{N \, f_x(x_i)\Delta V} \tag{2.c}$$

If we set the volume ΔV such that the data point x_i is inside a small hypersphere of volume ΔV and the following approximation is valid,

$$f_x(x_i)\Delta V \approx \int_{x \ \Delta V} f_x(x)dx \tag{3.a}$$

then, the right-hand side of eq.(3.a) is the probability that a sample is found in the volume ΔV, denoted by Prob{x in ΔV }. Note that it can be approximated by,

$$\text{Prob}\{x \text{ in } \Delta V\} = \frac{\text{number of samples in } \Delta V}{\text{total number of samples}} \tag{3.b}$$

Since the total number of samples is N, the denominator of (2.c) is written as,

$$Nf_X(x_i)\Delta V \approx N \,\text{Prob}\{x \text{ in } \Delta V\} \approx \text{number of samples found in } \Delta V \tag{3.c}$$

which can be obtained by counting the number of samples found in the hypersphere. In implementation, the counting in eq.(3.c) for each x_i can be done efficiently by finding first a set of hyperspheres which cover all given data samples in the feature space as in Fig. 2 and, then counting the data points inside the hyperspheres.

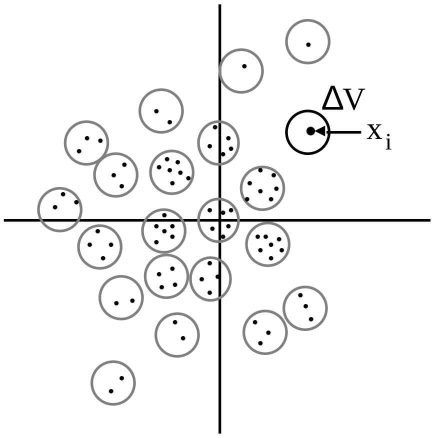

Figure 2. Computation of weights using clustering; Clustering is performed to find a set of hyperspheres covering the data samples.

This is easily accomplished by a few iterations of simple unsupervised clustering procedure in which a new cluster is generated if the distance to a nearest cluster exceeds a certain threshold. The threshold value is set up in such a way that each hypersphere corresponds to a cluster, and inside the hyperspheres, the probability density function, $f_x(x)$ does not change much so that the approximation in eq.(3.a) is valid. After a few iterations of clustering, the denominator in eq.(2.c) is computed as the counted value of samples found in the cluster which the i-th sample x_i belongs to. The numerator is calculated simply by multiplying the known PDF $f_x(x_i|C_1)$ with the volume ΔV and N_1 estimate.

B. Estimation of the Number of Samples Belonging to the Class of Interest

Due to limited prior knowledge, an accurate estimation of N_1 is another difficult task. However, the objective of the estimation here is to obtain a simple and reasonable estimate which can produce a meaningful initial cluster definition rather than pursuing a very accurate estimation.

The simplest method is counting the number of samples accepted by a given significance level. Define $N(\alpha)$ as the number of samples accepted by the significance testing with a significance level α, then, N(α) = (1-α) N_1 + N_{others}. (1-α) is the acceptance probability and N_{others} denotes the number of accepted samples that belong to the others class. In general, there is no guarantee that N_{others} is close to zero, or insignificant compared to (1-α)N_1. Nevertheless, the proposed method uses the simplest estimate of N_1, computed as,

$$N_1 = N(\alpha) / (1-\alpha) \tag{4}$$

ignoring N_{others}. This estimate always produces an over-estimated value, and the degree of over-estimation is significant when there is insufficient separability between the class of interest and the class of the others. It may be beneficial to use a large significance level of α hoping that a smaller acceptance probability (that is, tighter threshold) may exclude more samples of the others class. Note that an appropriate level of α is a priori unknown. In developing the initial clusters specification, however, experimental results show that this over-estimation is not critical to the performance, but an under-estimated value could be problematic since it causes non-trivial $\overline{w}_{i1}$ values and allows clusters generated in the region where most of the class-of-interest samples are located. These extraneous clusters would take a significant portions of class C_1 samples away.

C. Initial Cluster Definition

Once the weight factors are computed for all data samples in **X**, an unsupervised clustering is performed with the weights to find the clusters corresponding to C_{others}. For each cluster k corresponding to C_{others}, (that is, *k* = 2, ---, *L*), the cluster centroid is computed as the *effective* cluster mean,

$$M_k = \frac{1}{N_k} \sum_{i \quad I_k} \overline{w_{i1}} x_i \tag{5.a}$$

where I_k is the index set of the *k*-th cluster (i.e., if i $\quad I_k$, then $x_i \quad C_k$). N_k is the *effective* number of samples in the cluster and computed as,

$$N_k = \sum_{i \in I_k} \overline{w_{i1}} \qquad (5.b)$$

Note that the influence of data point x_i on the cluster means and number of samples is accordingly weighted by $\overline{w}_{i1}$. If second order statistics are necessary for clustering, then, the *effective* cluster covariance can also be computed with weights in a similar fashion. After each iteration of clustering, any cluster with a negligible effective number of members is deleted since most of the samples are from C_1. In the deletion, the ratio of the effective number to the actual sample number assigned to the cluster,

$$R_k = \frac{N_k}{\text{Number of samples in cluster } C_k} \qquad (6)$$

is also checked and any cluster with a small value of this ratio is deleted since most samples in the cluster have very negligible weight factors. When the number of class-of-interest samples, N_1, is under-estimated, this ratio checking is very important since extraneous clusters are generated in the region where most of the class-of-interest samples are located. This ratio checking should be also effective when the class-of-interest samples are distributed slightly differently from the known distribution function in some hyperspheres so that the numbers computed with eq.(2.a) deviate from those statistically expected. Without the ratio-checking, weights larger than they should be in some hyperspheres permit generating clusters of C_{others} which would take up significant portion of class-of-interest samples.

A few iterations of this unsupervised clustering with weights will suffice to provide a list of clusters corresponding to C_{others} and their initial specifications for the subsequent supervised clustering process.

D. Development of Class Statistics and Classification

Once the number of clusters and the specifications of the clusters are obtained through unsupervised clustering with weights, a conventional supervised clustering procedure can be started to refine the class statistics. The class statistics developed are used in the selected relative classification scheme. In certain cases, especially in analyzing high dimensional feature vectors, second order statistics, which are usually characterized by interband correlation structures, provide very crucial information to use in classification or in clustering [9]. In this case, a conventional clustering procedure such as the ISODATA [10] algorithm is not likely to perform well in developing class statistics since the algorithm does not account for interband correlation. In this case, a clustering method based on the *EM* algorithm [9] can be used. That is, in the m$\underline{\text{th}}$ iteration of clustering, weight factor, $w_{ik}[\hat{\Psi}^{(m)}]$, for *i* = 1,---, *N* and *k* = 1, ---, *L*, is computed as,

$$w_{ik}[\hat{\Psi}^{(m)}] = \frac{\hat{\pi}_k^{(m)} \hat{f}_x^{(m)}(x_i|C_k)}{\sum_{j=1}^{L} \hat{\pi}_j^{(m)} \hat{f}_x^{(m)}(x_i|C_j)} \tag{7}$$

where $\hat{f}_x^{(m)}(x_i \mid C_1) = f_X(x_i \mid C_1)$ for all m, and Ψ is the set of parameters of the unknown probability density functions(*E*xpectation - step). For example, if the unknown probability density functions are Gaussian, then $\Psi \equiv [\pi_2, ---, \pi_L, M_2, ---, M_L, \Sigma_2, ---, \Sigma_L]$. With the weight in eq.(7), a new maximum likelihood estimate of Ψ, (*i.e.*, $\hat{\Psi}^{(m+1)}$) is obtained (*M*aximization - step). These two steps are iteratively performed until convergence. Each iteration of these two steps is known to increase the joint likelihood of data samples [8]. After convergence, the estimates of Ψ specify the probability density functions of the clusters which can be used in the subsequent relative classification.

IV. EXPERIMENTS AND DISCUSSION

To test the performance of the proposed method, experiments are carried out with both simulated and real data. For simulated data, several bivariate Gaussian data sets are generated with different degrees of class separability. As real data, LANDSAT Thematic Mapper (TM) data are used. For comparison purposes, two additional classifiers are also used; the (fully supervised) maximum likelihood classifier (denoted as "REL-ML") designed with the known class statistics and the one based on the significance testing (denoted as "ABS-SIG"). The fully supervised maximum likelihood classifier is just to provide a point of comparison with the proposed classifier, that is, the lower bound of classification error which is ever achievable by the proposed method. On the other hand, the significance testing provides the performance result obtainable by a purely absolute scheme. In the significance testing, the best significance level is selected manually by testing the significance level in the interval [0.01, 0.99] in steps of 0.01. Therefore its results are the best ones obtainable by the significance testing.

Experiments with Simulated Data

For a test with simulated data, 1000 bivariate (q=2) Gaussian samples are generated for the class of interest, C_{int}, with zero mean and an identity covariance matrix. In the same way, 2000 Gaussian samples are generated for the others class, C_{others}, having the mean $[d,0]^T$, $d > 0$, and an identity covariance matrix. Therefore, N=3000 and $N_1 = 1000$.

With this set-up, the exact amount of overlap between two distributions can be calculated as (the "overlap" is defined here as the volume shared by the two probability density functions),

$$\text{Overlap}(d) = 1 - \frac{2}{\sqrt{2\pi}} \int_0^{d/2} \exp(-\frac{1}{2} s^2)\, ds$$

By varying the distance d between the two class means, data sets with different degrees of overlap can be simulated. In the simulation, d is increased from 0.1 to 5 in steps of 0.1. The value d = 0.1 simulates the case of 96.02% overlap between the two distributions, and d=5 presents an example of only 1.24 % overlap. To avoid any random errors due to the data generation process and its effect on evaluating experimental result, data sets are generated 50 times with different seed numbers and the averaged result is used in comparison.

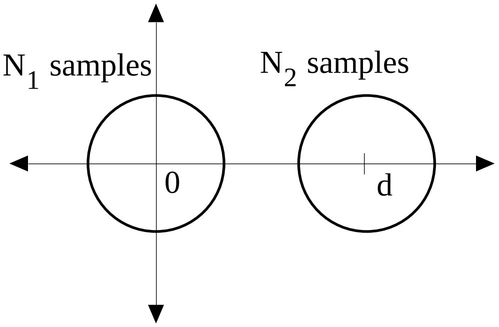

Figure 3. Simulated 2 class, 2 dimensional Gaussian data sets. C_{int}: 1000 samples with zero mean, C_{others}: 2000 samples with mean $[d,0]^T$. Both have an identity covariance matrix. (N_1 = 1000, N_2 = 2000, q=2).

A. Comparison of Classification Errors

Eq.(4) is used to obtain the N_1 estimate with varying significance level. The estimated number is shown in Fig. 4.

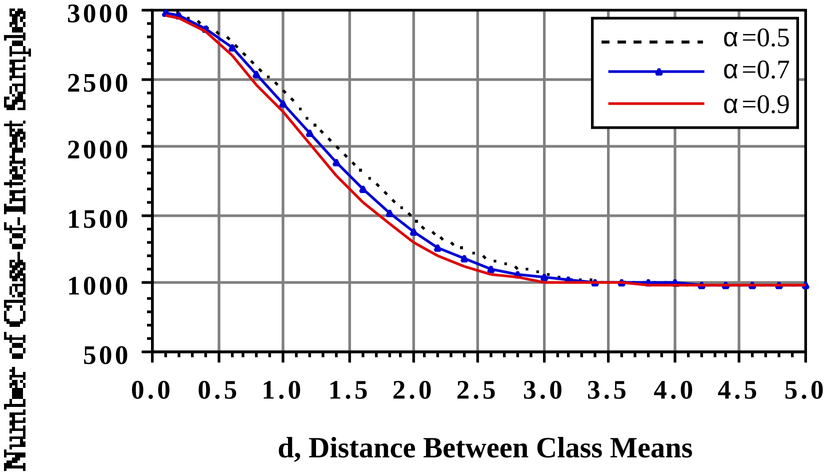

Figure 4. The estimated number of class-of-interest samples with different values of significance level α's in eq.(4).

The estimated values are observed to be not very different for different significance level α's especially when the classes are well separated (*i.e.*, d > 3, which corresponds to 13.36% overlap). When d < 3, there is significant degree of over-estimation.

Using the N_1 estimate, the weights $\overline{w}_{i1}$'s are computed and used in the unsupervised clustering to develop clusters corresponding to the others class. Any cluster having a negligible effective number of eq.(5.b) or a negligible ratio of eq.(6) is deleted. Without the ratio checking, the non-trivial weights $\overline{w}_{i1}$'s due to an under-estimated value of N_1 in the regions where the weights should be negligible would result in extraneous clusters and would cause large omission error. For those clusters, the effective numbers of samples would be much smaller than the actual sample numbers grouped to those clusters since significant portions of the samples in those clusters are from the class of interest. When the actual distribution of the class-of-interest samples is slightly different from that predicted by the probability density function $f_x(x|C_1)$, those extraneous clusters can also be observed even though N_1 is not greatly under-estimated.

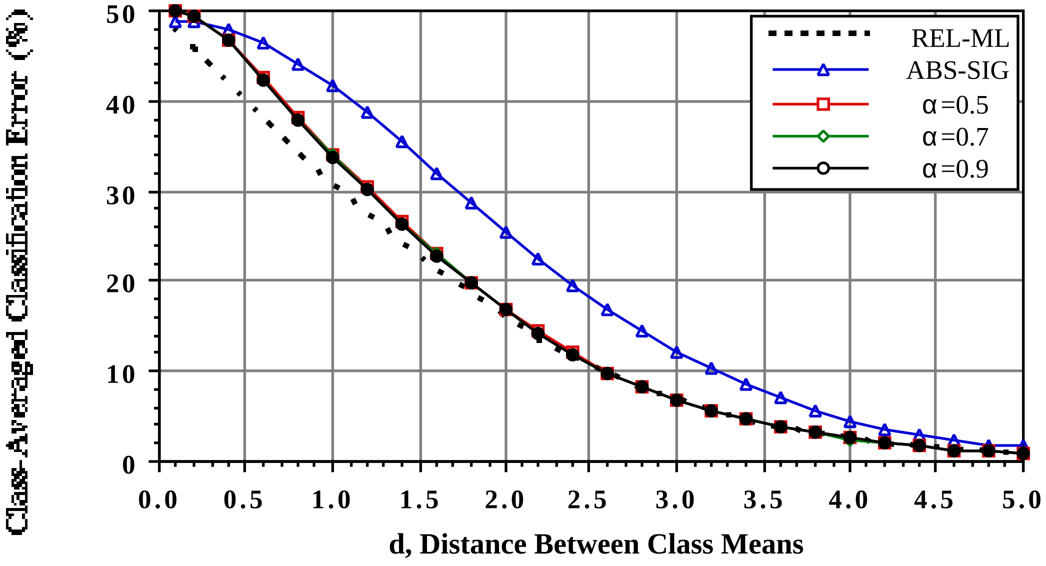

Figure 5. Class-averaged classification error comparison. The proposed method is denoted by the α value used in estimating the number of class-of-interest samples with eq.(4). "REL-ML" is the relative ML classifier with known class statistics and "ABS-SIG" is the best attainable result by significance testing.

Figure 5 compares the class-averaged classification errors of the relative maximum likelihood classifier (REL-ML), the classifier based on significance testing (ABS-SIG), and the proposed classifier with three different significance level α's for the N_1 estimation (denoted with three different α values). The class-averaged classification error is a simple average of the omission and commission errors.

omission error $\varepsilon_0 = P$ {x is decided as C_{others} | x C_{int}}
commission error $\varepsilon_1 = P$ {x is decided as C_{int} | x C_{others}}

While the significance testing results in about 5 ~ 12 % greater error than the relative maximum likelihood classifier unless d is sufficiently large, the proposed method closely follows the performance of the maximum likelihood classifier. Only when the overlap between two classes is significant (for example, see the case d < 2, 31.14% overlap), there is some error increase compared to the maximum likelihood classifier, but the deviation is at most less than 5 %.

B. Sensitivity to N_1 Estimate

To analyze the reason for the performance deviation for d<2.0 in Fig.5 and the sensitivity of the proposed classifier to the N_1 estimate in computing the weights $\overline{w}_{i1}$'s, several different values of N_1 are used instead of the estimated values and its classification result is analyzed as in Fig. 6.

There is almost negligible difference in class-averaged classification error when N_1 is varied from 750 to 1500 (not shown). When an over-estimated N_1 is used, there is as much as 2% (N_1 = 2000, 100% over-estimation) or 5% (N_1 = 3000, 200% over-estimation) error increase compared to the maximum likelihood classifier for d < 2. The observation that the estimation method of N_1 always over-estimates as seen in Fig.4 for d < 2.5 and severe over-estimation brings maximum 5% of classification error increase over the REL-ML as seen in Fig.6 explains the reason for the increased deviation for d < 2 in Fig.5.

One important observation is that even though the simple method in eq.(4) gives very rough and over-estimated N_1 and in some cases the over-estimated N_1 increases the commission error, but its effect on classification accuracy is much less in the proposed classifier compared to the significance testing in which improper significance level results in classification accuracy degradation of 5% ~ 50% depending on distribution overlap. Therefore, the proposed method can be said relatively insensitive to the significance level (which is used for N_1 estimation).

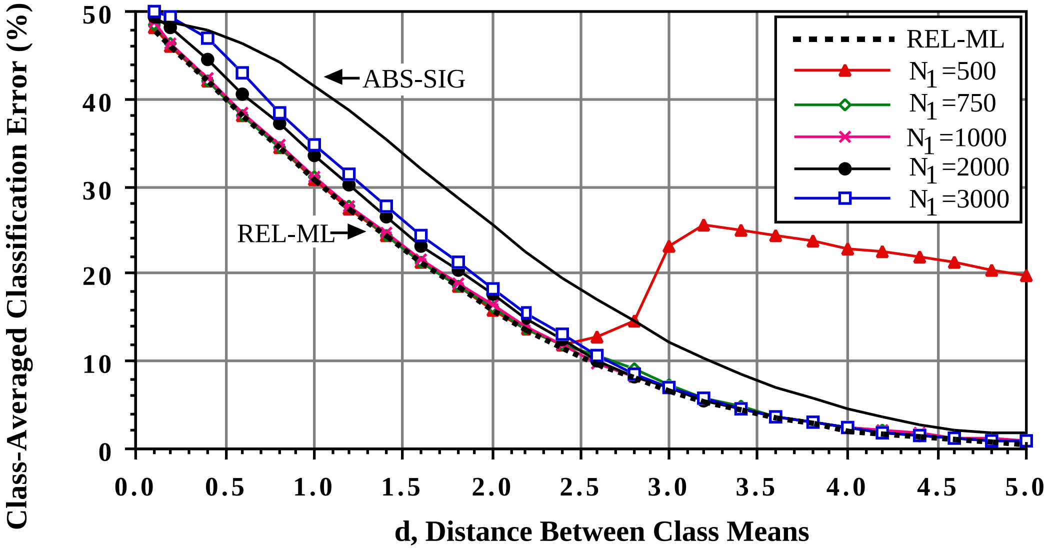

Figure 6. Sensitivity of the proposed method to the estimate N_1. Several different values of N_1 are used in computing the weights $\overline{w}_{i1}$'s without estimating (the true value of N_1 is 1000).

The proposed method is also observed very tolerable on the degree of over-estimation, however, it is less so on under-estimation as shown in Fig. 6. See the case N_1=500 (50% under-estimation); when d > 2.5, the class-averaged classification error increases since the clusters containing non-trivial portion of the class-of-interest samples survive the cluster deletion test of eq.(5.b, 6) and many class-of-interest samples are deleted to increase omission error. Note that the N_1 estimate with eq.(4) is in general slightly over-estimated as shown in Fig. 4 due to the commission of "others" samples. The under-estimation is not so problematic in reality unless the training samples of the class of interest are not representative enough to adequately model its distribution function.

Experiment with Thematic Mapper Data

A real data test is carried out using the LANDSAT Thematic Mapper data acquired over an agricultural area in Tippecanoe County, Indiana in July, 1986. All seven features are used in the classification. From the ground truth data, four different information classes are identified as in Table 1 and about 10% of the samples are randomly selected from each class to serve as training samples. In the test of the proposed classifier using the real data, each information class is assumed to be the class of interest one by one (indicated in the header of column 2 to 6 in Table 2) and the other three as the class of the others. Therefore, the test is still a two-way classification problem. The information classes are modeled by several sub-classes each of which has the multivariate Gaussian PDF. To obtain a set of constituent sub-classes as in Table 1, clustering is performed first on the selected training samples belonging to each information class. Then, the training samples clustered to each sub-class are used to calculate the mean and covariance of its Gaussian PDF.

Table 1. Training and test samples of LANDSAT Thematic Mapper data

Information Classes	Number of Sub-Classes	Number of Samples	
		Training	Test
Corn	2	913	9371
Soybeans	2	824	8455
Wheat	4	181	1923
Alfalfa/Oats	4	206	2175
Total	12	2124	21924

In classification, the whole data set is first divided by the maximum likelihood classifier into n sub-groups where n is the number of sub-classes of a given information class as in Table 1. For each sub-group, the proposed method is applied to identify the samples belonging to the corresponding sub-class. As before, the maximum likelihood classifier (REL-ML) designed with the total 12 sub-classes and the significance testing with the manually selected best significance level are used for comparison. The test result is summarized in Table 2.

Table 2. Comparisons of Classification Error in %.

Error Criterion	Classifier	Corn	Soybean	Wheat	Alfafa/Oats
Omission	REL-ML	3.04	13.65	13.36	23.95
Error	ABS-SIG	4.64	12.94	10.45	24.14
ε_0	Proposed	3.69	6.02	10.66	18.53
Commission	REL-ML	1.15	3.18	2.04	6.89
Error	ABS-SIG	2.79	11.17	5.38	21.31
ε_1	Proposed	1.68	7.02	3.6	16.38
Class Averaged	REL-ML	2.1	8.42	7.7	15.42
Error	ABS-SIG	3.72	12.05	7.92	22.72
$(\varepsilon_0+\varepsilon_1)/2$	Proposed	2.69	6.52	7.13	17.45
Total	REL-ML	1.95	7.19	3.02	8.57
Error	ABS-SIG	3.58	11.85	5.82	21.59
$\pi_1\varepsilon_0 + (1-\pi_1)\varepsilon_1$	Proposed	2.53	6.64	4.21	16.59

The relative class separability can be predicted in some degree by checking the commission error ε_1 of the significance testing; the class of corn and wheat have relatively small commission errors, therefore each distribution of two classes is seen relatively separable from others. For these two classes, the classification errors of all three classifier are almost equivalent. However, the separability of the class of alfafa/oats from the others is not so large since the significance testing results in 21% of the commission error. In this case, the proposed method's classification error is only 2% (class-averaged classification error sense) and 8% (total classification error sense) higher than the maximum-likelihood classifier result, while the significance testing classification error is 7% and 13% higher, respectively, than the maximum likelihood classification result.

When estimating N_1, five different values of α (0.9, 0.8, 0.7, 0.6 and 0.5) are used to observe that the estimated numbers N_1 are mostly over-estimated and that there are less than 1% of the differences in the classification error even though there are large differences in the degree of over-estimation (21% ~ 177%). Table 2 shows only the result with α = 0.5. The proposed method is seen to perform better in all classes by about 1 ~ 6 % than the *best* significance testing case where the significance levels are strenuously chosen manually.

Compared to the fully supervised maximum likelihood classifier which requires a complete list of classes with their class statistics, the proposed method achieves comparable classification performance even though prior knowledge is provided only for the specific information class under consideration. The computational complexity increases over the relative maximum likelihood classifier, but not prohibitively so in view of the time savings for the manual portion of the analysis task. In the experiment with Thematic Mapper data in identifying one information class, it takes about 3 times more computational time than the maximum likelihood classifier.

V. CONCLUDING REMARKS

In this paper, we have proposed a new partially supervised classification method using unsupervised clustering. Since the definition and statistics of the "others" class are automatically developed through a weighted unsupervised clustering procedure, the user needs to supply prior information for a particular class one actually wants to identify. This operational simplicity should make this method useful in many practical applications.

Experiments with simulated and real Thematic Mapper data show that the proposed method is definitely better than the conventional approach using the significance testing even if the best optimal significance level is manually provided. On the contrary to the large range of classification performance variation due to significance level input by the user in case of the significance testing, the proposed method is much less sensitive to the significance level provided by the user for estimating N_1.

It is also pretty much comparable to the fully supervised maximum likelihood classifier unless the overlap of class distribution is very significant. The classification accuracy degradation when the class distribution is heavily overlapped is found to be caused by the over-estimated number of samples belonging to the class of interest. Since a relatively accurate estimate achieves classification performance very close to that of the fully supervised maximum-likelihood classifier, a better estimation method of N_1 deserves further investigation.

REFERENCES

[1] K. Fukunaga, R. R. Hayes, and L. M. Novak, "The Acquisition Probability for a Minimum Distance One-class Classifier," IEEE Trans. on Aerospace and Electronic Systems, AES-23, pp.493-499, 1987.

[2] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd edition, Academic Press, New York, 1990.

[3] T. F. Quatieri, "Object Detection by two-dimensional linear prediction," Proc. of IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp.108-111, Apr., 1983.

[4] P. H. Swain and S. Davis, Remote Sensing - The Quantitative Approach, McGraw-Hill Book Company, New York, 1978.

[5] B. M. Shahshahani and D. A. Landgrebe, "The Effect of Unlabeled Samples in Reducing the Small Sample Size Problem and Mitigating the Hughes Phenomenon," IEEE Trans. on Geoscience and Remote Sensing, Vol.32, No.5, pp.1087-1095, Sept. 1994.

[6] C. W. Therrien, T. F. Quatieri, and D. E. Dudgeon, "Statistical Model-Based Algorithms for Image Analysis," Proc. of IEEE, Vol.74, No.4, pp.532-551, April, 1986.

[7] B. Jeon and D. A. Landgrebe, "A New Supervised Absolute Classifier," Proc. of IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp.2363-2366, 1990.

[8] M. Titterington, *et al.*, Statistical Analysis of Finite Mixture Distributions, John Wiley & Sons, New York, 1985.

[9] C. Lee and D. A. Landgrebe, "Analyzing High Dimensional Multispectral Data," IEEE Trans. on Geoscience and Remote Sensing, Vol.31, No.4, pp.792-800, July 1993.

[10] D. J. Hall and G. B. Ball, "ISODATA : A novel method of data analysis and pattern classification," Technical Report, Stanford Research Institute, Menlo Park, CA, 1965.

Byeungwoo Jeon

Byeungwoo Jeon(S'88-M'92) received the B.S. degree (magna cum laude)in 1985, the M.S. degree in 1987 from the Department of Electronics Engineering, Seoul National University, Korea, and the Ph.D. degree from the School of Electrical Engineering, Purdue University, West Lafayette, IN, in 1992. From 1993 to 1997, he was with the Signal Processing Laboratory, Samsung Electronics, Korea, where he was involved in video compression, the development of digital broadcasting satellite receiver, and other MPEG-related multimedia applications. Since September 1997, he has been with the School of ECE, Sungkyunkwan University, Korea, as an assistant professor. His research interests include multimedia signal processing, image compression, statistical pattern recognition, and remote sensing. Dr. Jeon is a member of Tau Beta Pi and Eta Kappa Nu.

David A. Landgrebe

Dr. Landgrebe holds the BSEE, MSEE, and PhD degrees from Purdue University. He is presently Professor of Electrical and Computer Engineering at Purdue University. His area of specialty in research is communication science and signal processing, especially as applied to Earth observational remote sensing. He was President of the IEEE Geoscience and Remote Sensing Society for 1986 and 1987 and a member of its Administrative Committee from 1979 to 1990. He received that Society’s Outstanding Service Award in 1988. He is a co-author of the text, *Remote Sensing: The Quantitative Approach*, and a contributor to the book, *Remote Sensing of Environment*, and the *ASP Manual of Remote Sensing (1st edition)*. He has been a member of the editorial board of the journal, *Remote Sensing of Environment*, since its inception.

Dr. Landgrebe is a Life Fellow of the Institute of Electrical and Electronic Engineers, a Fellow of the American Society of Photogrammetry and Remote Sensing, a member of the Society of Photo-Optical Instrumentation Engineers and the American Society for Engineering Education, as well as Eta Kappa Nu, Tau Beta Pi, and Sigma Xi honor societies. He received the NASA Exceptional Scientific Achievement Medal in 1973 for his work in the field of machine analysis methods for remotely sensed Earth observational data. In 1976, on behalf of the Purdue's Laboratory for Applications of Remote Sensing which he directed, he accepted the William T. Pecora Award, presented by NASA and the U.S. Department of Interior. He was the 1990 individual recipient of the William T. Pecora Award for contributions to the field of remote sensing. He was the 1992 recipient of the IEEE Geoscience and Remote Sensing Society’s Distinguished Achievement Award.