Special Session on Hyperspectral Imaging

Maite Termenon[1]

[1]Computational Intelligence Group

2012 March 9

Article

IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, VOL. 37, NO. 2, MARCH 1999

Partially Supervised Classification Using Weighted Unsupervised Clustering

Byeungwoo Jeon, *Member, IEEE,* and David A. Landgrebe, *Life Fellow, IEEE*

Manuscript received April 9, 1998; revised August 28, 1998.

B. Jeon is with the School of Electrical and Computer Engineering, Sung Kyun Kwan University, Suwon, Korea.

D. A. Landgrebe is with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907-1285, USA (e-mail: landgreb@ecn.purdue.edu

Publisher Item Identifier S 0196-2892(99)01980-4.

Resume

Keywords: One-class classifier, partially supervised classifier, significance testing, single hypothesis testing, unsupervised clustering.

Problem Definition:

- **Prior** information available only for the class of interest: **PDF of class of interest**
- **Key** Problem: how to define and find statistics for the other clusters.
- 3 Steps:
  - To assign weight factors to each sample.
  - To define the initial others class clusters.
  - Iteratively, refine clusters statistics and **ML classifier** makes decisions using clusters statistics.
- **Applications**: target detection, object detection out of various backgrounds, texture detection, cloud identification.

Assumptions:

- C_{int} is modeled by a known PDF with zero Mean and identity covariance matrix.

Experiments:

- Simulated data:
  - Bivariate Gaussian data sets are Generated with different degrees of Class separability.
- Real data: LANDSAT Thematic Mapper
- Algorithms to compare:
  - REL-ML (fully supervised)
  - One based on the significance testing (ABS-SIG)

Outline

Outline

Definition

- Given data set: $\mathbf{X} \equiv \{x_1, \ldots, x_N\}$, where q-dimensional feature vector, x_i.
- N_1: *unknown* number of samples that belong to the class of interest, C_{int}.
- C_{others}: might consist of several subclasses none of which are of interest.
- C_{int} is assumed to be modeled by a known probability density function (PDF), $f_x(x \mid C_{int})$.

Problem

- The problem is considered as an unsupervised clustering problem with initially one known cluster.
- Solution:
  1. Each data sample is assigned a weight factor indicating likelihood of being from "the others" class.
  2. To develop the initial definition of clusters corresponding to the others class using the weighted unsupervised clustering.
  3. Cluster statistics are iteratively refined, and a conventional relative classifier such as the maximum likelihood (ML) classifier makes decisions using the cluster statistics.

Outline

Weight Factors

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

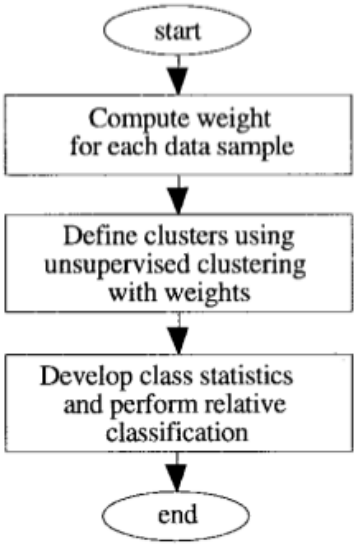


Fig. 1. Flowchart of proposed partially supervised classification method.

Computation of Weights

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

$$f_x(x) = \sum_{k=1}^{L} \pi_k f_x(x|C_k) \tag{1}$$

where π_k and $f_x(x|C_k)$ are respectively the prior probability and probability density function of the kth class, $k = 1, \cdots, L$, and $\pi_1 + \cdots + \pi_L = 1$. The notation of C_1 and $C_2, \cdots, C_L$ means that $C_1 = C_{\text{int}}$ and $C_2, \cdots, C_L$ are the subclasses of C_{others}. According to the assumption, only $f_x(x|C_1)$ is known *a priori*. We develop the probability distribution functions

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 (2a) which is the relative likelihood of not belonging to C_{int}

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

where

$$w_{i1} = \pi_1 \frac{f_x(x_i|C_1)}{f_x(x_i)} = \frac{N_1 f_x(x_i|C_1)}{N f_x(x_i)}. \tag{2b}$$

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 $\boldsymbol{X}$ 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 nonparametric density estimation would require complex computation, a practical approximation is made by noting that w_{i1} can be expressed as a ratio

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

Computation of Weights

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

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

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

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

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

$$\begin{aligned} N f_x(x_i)\Delta V &\approx N \text{ Prob}\{x \text{ in } \Delta V\} \\ &\approx \text{number of samples found in } \Delta V \end{aligned} \tag{3c}$$

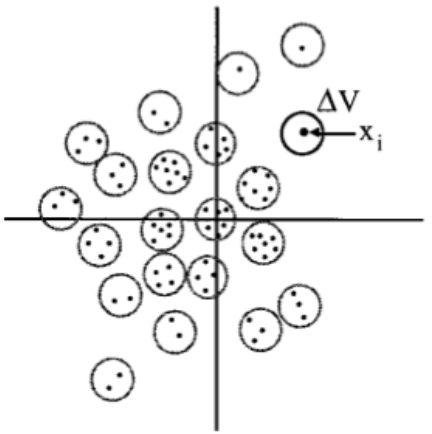


Fig. 2. Computation of weights using clustering; clustering is performed to find a set of hyperspheres covering the data samples.

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

- An accurate estimation of N_1 is another difficult task.
- Objective: to obtain a simple and reasonable estimate which can produce a meaningful initial cluster definition.
- Simplest method: counting the number of samples accepted by a given significance level, α.
- $N(\alpha) = (1-\alpha)\,N_1 + N_{others}$

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-

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

- This estimate always produces an over-estimated value.
- Degree of over-estimation is significant when there is insufficient separability between C_{int} and C_{others}.
- An appropriate level of α is a priori unknown, but it is recomended to use a large significance level of α hoping that a smaller acceptance probability may exclude more samples of the others class.
- Experimental results: over-estimation is not critical to the performance, but an under-estimated value could be problematic since it causes nontrivial $\bar{w}_{i1}$ values and allows clusters generated in the region where most of the class-of-interest samples are located.

Initial Clusters Definition

Once the weight factors are computed for all data samples in $\boldsymbol{X}$, 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, $k = 2, \cdots, L$), the cluster centroid is computed as the *effective* cluster mean:

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

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

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

Note that the influence of data point x_i on the cluster means and number of samples is accordingly weighted by $\overline{w}_{i1}$.

Datasets for training and testing

- 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.

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

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

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

- 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.

Development of Class Statistics and Classification

- Once the number of clusters and specifications of clusters are obtained, a conventional supervised clustering procedure can be started to refine the class statistics.

In this case, a clustering method based on the EM algorithm [9] can be used. That is, in the mth iteration of clustering, weight factor, $w_{ik}[\hat{\Psi}^{(m)}]$, for $i = 1, \cdots, N$ and $k = 1, \cdots, L$, is computed as

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

where $\hat{f}_x^{(m)}(x_i|C_1) = f_x(x_i|C_1)$ for all m, and Ψ is the set of parameters of the unknown probability density functions (Expectation-step). With the weight in (7), a new maximum likelihood estimate of Ψ, (i.e., $\hat{\Psi}^{(m+1)}$) is obtained (Maximization-step).

- After convergence, the estimates of Ψ specify the probability density functions of the clusters which can be used in the subsequent relative classification.

Outline

Experiments

- Simulated data: several bivariate Gaussian data sets are generated with different degrees of class separability.
- Real data: LANDSAT Thematic Mapper (TM) data.
- Comparison purposes:
  - Fully supervised maximum likelihood classifier ("REL-ML"): to provide the lower bound of classification error which is ever achievable by the proposed method.
  - One based on the significance testing ("ABS-SIG"): to provide the performance result obtainable by a purely absolute scheme (best significance level is selected manually by testing the significance level in the interval [0.01, 0.99] in steps of 0.01).

Simulated Data

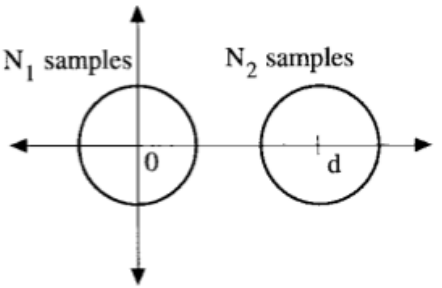


Fig. 3. Simulated 2 class, 2-D Gaussian data sets. C_{int}: 1000 samples with zero mean, C_{others}: 2000 samples with mean $[d, 0]^T$. Both have an identity covariance matrix. ($N_1 = 1000, N_2 = 2000, q = 2$.)

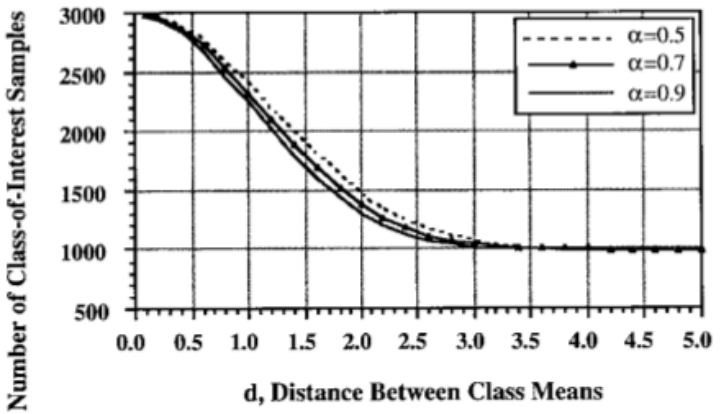
With this setup, the exact amount of overlap between two distributions can be calculated as (the "overlap" is defined here as the volume shared by the two probability density functions),

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

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

Comparison of Classification Errors

- Equation (4) is used to obtain the N_1 estimate varying significance level.
- Using the N_1 estimate, weights $\overline{w}_{i1}$ are computed and used in the unsupervised clustering to develop clusters corresponding to the others class. Some clusters are deleted taking into account (5b) and (6).



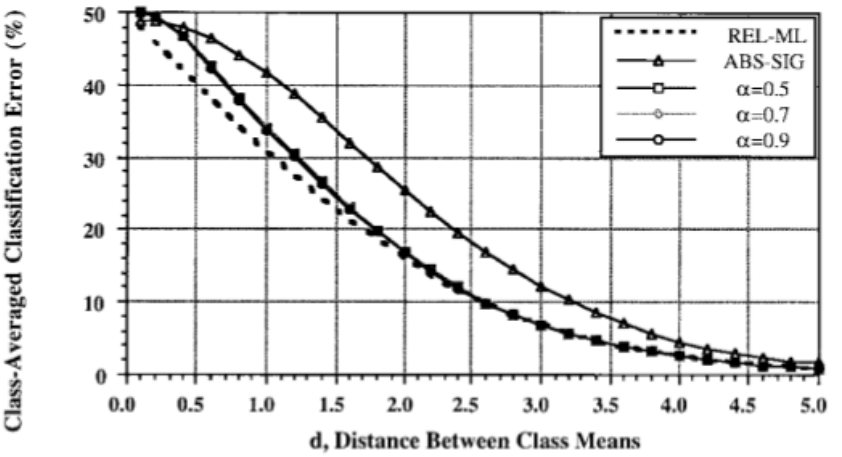


Fig. 5. Class-averaged classification error comparison. The proposed method is denoted by the α value used in estimating the number of class-of-interest samples with (4). "REL-ML" is the relative ML classifier with known class

Sensitivity to N1 Estimate

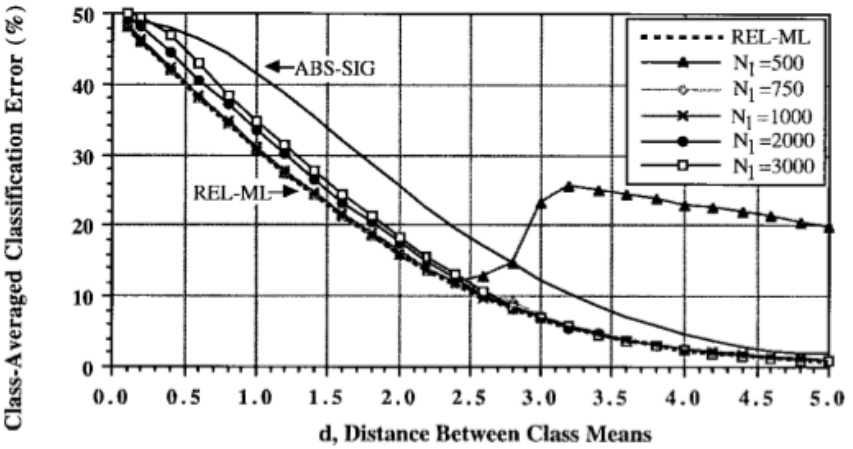


Fig. 6. Sensitivity of the proposed method to the estimate N_1. Several different values of N_1 are used in computing the weights $\overline{w}_{i1}$'s without estimating (the true value of N_1 is 1000).

- proposed method can be said relatively insensitive to the significance level.
- proposed method is also observed very tolerable on the degree of over-estimation, however, it is less so on under-estimation

Real Data

- From the ground truth data, four different information classes are identified.
- About 10% of the samples are randomly selected from each class to serve as training samples.
- In the test, each information class is assumed to be the C_{int} one by one and the other three as the C_{others}.
- Information classes are modeled by several subclasses each of which has the multivariate Gaussian PDF.

Real Data

- Clustering is performed first on the selected training samples belonging to each information class.
- Training samples clustered to each subclass are used to calculate the mean and covariance of its Gaussian PDF.
- In classification, the whole data set is first divided by the ML classifier into n subgroups where n is the number of subclasses of a given information class.
- For each subgroup, the proposed method is applied to identify the samples belonging to the corresponding subclass.

Real Data

TABLE I
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

TABLE II
COMPARISONS OF CLASSIFICATION ERROR IN PERCENT

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