Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 41 (2008) 2842–2854
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

A study of regularized Gaussian classifier in high-dimension small sample set case based on MDL principle with application to spectrum recognition

Ping Guo[a,b,*], Yunde Jia[a], Michael R. Lyu[c]

[a]School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, PR China
[b]Laboratory of Image Processing and Pattern Recognition, Beijing Normal University, Beijing, 100875, PR China
[c]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, PR China

Received 6 December 2005; received in revised form 24 January 2008; accepted 2 February 2008

Abstract

In classifying high-dimensional patterns such as stellar spectra by a Gaussian classifier, the covariance matrix estimated with a small-number sample set becomes unstable, leading to degraded classification accuracy. In this paper, we investigate the covariance matrix estimation problem for small-number samples with high dimension setting based on minimum description length (MDL) principle. A new covariance matrix estimator is developed, and a formula for fast estimation of regularization parameters is derived. Experiments on spectrum pattern recognition are conducted to investigate the classification accuracy with the developed covariance matrix estimator. Higher classification accuracy results are obtained and demonstrated in our new approach.

Keywords: Classification; Covariance matrix estimation; Discriminant analysis method; Regularization parameter selection; Minimum description length

1. Introduction

Spectrum recognition has a wide range of applications, such as chemical element identification, stellar classification, and matter structure analysis. For spectral data, the number of variables (wavelengths) is much higher than that of training samples; therefore, spectral data are severely ill-posed. Due to such high dimensionality, the common multivariate classification methods of linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) cannot be directly applied because of the matrix singularity problem [1].

Spectrum recognition is usually a high-dimensional small sample set classification problem. Generally speaking, classification has two aspects: supervised classification (discrimination or simply classification) and unsupervised classification (clustering). In recent years, several classification algorithms have been developed to partition a data set into pre-defined classes. When the data are viewed as arising from two or more clusters mixed in varying proportions, we can use the finite Gaussian mixture distribution to analyze the data set. The Gaussian mixture distribution analysis method has been employed widely in a variety of important practical situations, where the likelihood approach to the fitting of Gaussian mixture models has been utilized extensively [2–5].

When classifying data with the Gaussian mixture model, the mean vector and covariance matrix of each component are not known in advance, and they have to be estimated from the given data set. While a large-size data set is desirable for estimating the parameters more accurately, in the real world, often only a small-size data set can be obtained because of some restriction, e.g., high cost in collecting large-size data sets. For a relatively small-number sample data set, if the dimension d of variable $\mathbf{x}$ is comparable to the number of training samples n_j in class j, the problem may become poorly posed. Worse, if the number n_j of training samples is less than the dimensionality, the problem becomes ill-posed. In this case, not all parameters can be properly estimated, and the classification accuracy is degraded.

* Corresponding author at: School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, PR China. Tel.: +86 10 58806662; fax: +86 10 58809444.
E-mail addresses: pguo@ieee.org (P. Guo), lyu@cse.cuhk.edu.hk (M.R. Lyu).

0031-3203/$30.00 © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.02.004

There are two possible solutions for this kind of problem: one is dimensionality reduction [6,7], and the other is regularization [1,8]. Regularization is the procedure of allowing parameters bias toward what are thought to be more plausible values, which reduces the variance of the estimates at the cost of introducing bias. Besides the regularization techniques can be used to sparse nonparametric density estimation in high dimension case [9], the regularization techniques have been highly successful in classifying small-number data with some heuristic approximations [1,8,10,11]. However, these methods, such as regularized discriminant analysis (RDA) [10], require users to select regularization parameters (or called *model*) with some statistical techniques like leave-one-out cross-validation [11–14], which is computation-expensive. Furthermore, a recent study shows that cross-validation performance is not always good in the selection of linear models [15] in some cases. Therefore, it is worthy to further investigate this problem.

Originally proposed as an estimation criterion by Rissanen [16,17], the minimum description length (MDL) principle can be applied to universal coding, linear regression, and density estimation problems. *The central idea of this principle is to represent an entire class of probability distributions as models by a single "universal" representative model, such that it would be able to imitate the behavior of any model in the class. The best model class for a set of observed data is the one whose representative permits the shortest coding of the data. The MDL estimates of both the parameters and their total number are consistent; i.e., the estimates converge and the limit specifies the data generating model* [17]. The codelength[1] criterion of MDL involves in the Kullback–Leibler divergence [18,19]. MDL principle has a wide applications, such as clustering problem [20]. In this paper, based on the MDL principle with the mixture model analysis, we present the results of investigating covariance matrix estimation and regularization parameter selection in the Gaussian classifier for the small-sample set with high-dimension classification problem.

2. Theoretical background

2.1. Classification with finite Gaussian mixture model

In pattern recognition problem, we have a set of data samples, each consisting of measurements on a set of variables with associated labels, the class types. They are used as exemplars in the classifier design [21]. In clustering we need to estimate *prior* probability and *posterior* probability in the classifier design. If these probabilities are known, it becomes a classification problem. So clustering is more general than classification in the mixture model analysis case. Let us consider the general case first.

The data points $D = \{\mathbf{x}_i\}_{i=1}^{N}$ to be classified are assumed to be samples from a mixture of k Gaussian densities with joint probability density of which the mathematical expressions are shown as follows:

$$p(\mathbf{x}, \Theta) = \sum_{j=1}^{k} \alpha_j G(\mathbf{x}, \mathbf{m}_j, \mathbf{\Sigma}_j)$$

$$\text{with } \alpha_j \geqslant 0 \quad \text{and} \quad \sum_{j=1}^{k} \alpha_j = 1, \tag{1}$$

where

$$G(\mathbf{x}, \mathbf{m}_j, \mathbf{\Sigma}_j) = \frac{\exp[-\frac{1}{2}(\mathbf{x}-\mathbf{m}_j)^{\mathrm{T}}\mathbf{\Sigma}_j^{-1}(\mathbf{x}-\mathbf{m}_j)]}{(2\pi)^{d/2}|\mathbf{\Sigma}_j|^{1/2}} \tag{2}$$

is a general multivariate Gaussian density function, $\mathbf{x}$ denotes a random vector, d is the dimension of the $\mathbf{x}$, and parameter $\Theta = \{\alpha_j, \mathbf{m}_j, \mathbf{\Sigma}_j\}_{j=1}^{k}$ is a set of finite mixture model parameter vectors. Here α_j is the *prior* probability, $\mathbf{m}_j$ is the mean vector, and $\mathbf{\Sigma}_j$ is the covariance matrix of the jth component. Based on the given data set, these parameters can be estimated by the maximum likelihood (ML) method with expectation-maximum (EM) algorithm [22,23].

In the Gaussian mixture model case, the Bayesian decision rule is applied to classify the vector $\mathbf{x}$ into class j with the largest *posterior* probability. The *posterior* probability $P(j|\mathbf{x}, \Theta)$ represents the probability that the sample $\mathbf{x}$ belongs to class j. We use Bayesian decision $j^* = \arg\max_j P(j|\mathbf{x}, \Theta)$ to classify $\mathbf{x}$ into class j^*. The probability functions $P(j|\mathbf{x}, \Theta)$ are usually unknown and have to be estimated from the training samples. With the ML method estimated parameter $\widehat{\Theta}$, the *posterior* probability can be written in the form

$$P(j|\mathbf{x}, \widehat{\Theta}) = \frac{\widehat{\alpha}_j G(\mathbf{x}, \widehat{\mathbf{m}}_j, \widehat{\mathbf{\Sigma}}_j)}{p(\mathbf{x}, \widehat{\Theta})} \quad \text{with}$$
$$j = 1, 2, \ldots, k. \tag{3}$$

Taking the logarithm to the above equation and omitting the common factors of the classes, the classification rule becomes

$$j^* = \arg\min_j d_j(\mathbf{x}), \quad j = 1, 2, \ldots, k \tag{4}$$

with

$$d_j(\mathbf{x}) = (\mathbf{x}-\widehat{\mathbf{m}}_j)^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_j^{-1}(\mathbf{x}-\widehat{\mathbf{m}}_j) + \ln|\widehat{\mathbf{\Sigma}}_j| - 2\ln\widehat{\alpha}_j. \tag{5}$$

This equation is often called the discriminant function for the jth class in the literature [1]. Furthermore, if the *prior* probability $\widehat{\alpha}_j$ is the same for all classes, the term $2\ln\widehat{\alpha}_j$ can be omitted and the discriminant function reduces to a simpler form [24].

2.2. Covariance matrix estimation

When the sample number is small, the sample-based estimation of class-specific covariance matrix becomes inaccurate, resulting in lowered classification accuracy. To solve this problem, several techniques are proposed, such as LOOC as well as its extensions bLOOC1 and bLOOC2 [11–14]. LOOC was proposed by Hoffbeck and Landgrebe [11], who examine the diagonal sample covariance matrix, the diagonal common covariance matrix, and some pair-wise mixtures of those matrices.

[1] A term codelength is just another way to express a probability distribution or a model.