Available online at www.sciencedirect.com
SCIENCE DIRECT®

Pattern Recognition 39 (2006) 1346–1358

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Feature representation and discrimination based on Gaussian mixture model probability densities—Practices and algorithms

Pekka Paalanen, Joni-Kristian Kamarainen*, Jarmo Ilonen, Heikki Kälviäinen

Department of Information Technology, Lappeenranta University of Technology, P.O. Box 20, FI-53851 Lappeenranta, Finland

Received 15 July 2005; accepted 10 January 2006

Abstract

This study promotes the use of statistical methods in specific classification tasks since statistical methods have certain advantages which advocate their use in pattern recognition. One central problem in statistical methods is estimation of class conditional probability density functions based on examples in a training set. In this study maximum likelihood estimation methods for Gaussian mixture models are reviewed and discussed from a practical point of view. In addition, good practices for utilizing probability densities in feature classification and selection are discussed for Bayesian and, more importantly, for non-Bayesian tasks. As a result, the use of confidence information in the classification is proposed and a method for confidence estimation is presented. The propositions are tested experimentally.

Keywords: Gaussian mixture model; EM; Classifier; Confidence; Highest density region

1. Introduction

Recently, black box and gray box pattern recognition (PR) and feature classification methods have proved to be very powerful and methods such as multi-layer perceptron neural networks [1] and support vector machines [2] are frequently applied with a great success. Furthermore, other novel methods seem to embed feature selection into a classifier synthesis as, for example, in the AdaBoost boosting algorithm [3]. These powerful methods are also state-of-the-art methods in practice and it is justifiable to ask whether structural and statistical PR approaches are still relevant.

Drawbacks in black and gray box PR methods are often their incapability to provide confidence information for their decision or difficulty in incorporating risk and cost models into the recognition process. In many applications it is not sufficient just to assign one predefined class to new observations; for example, in face detection facial evidence, such as eye centers and nostrils, should be detected from a scene and provided to the next processing level in ranked order (best candidates first) in order to perform detection computationally efficiently [4,5]. Gray box methods may include the confidence information as an ad hoc solution. Statistical methods, on the other hand, usually provide the information in an interpretable form along with sufficient mathematical foundations. Statistical methods thus provide some advantages over black box methods; the decision making is based on an interpretable basis from which the most probable or lowest risk (expected cost) option can be chosen (e.g. Bayesian decision making [6]) and different observations can be compared based on their statistical properties.

In a typical PR problem, features from known observations, a training set, are provided and necessary statistics must be established for recognition of unknown observations and estimation of confidence. A class of patterns is typically represented as a probability density function (pdf) of features. Selection of proper features is a distinct and application specific problem, but as a more general consideration, finding a proper pdf estimate has a crucial impact on successful recognition. Typically, the form of the pdf is somehow restricted and the search is reduced to a problem of

* Corresponding author. Tel.: +358 5 6212844; fax: +358 5 6212899.
E-mail addresses: paalanen@lut.fi (P. Paalanen), jkamarai@lut.fi, Joni.Kamarainen@lut.fi (J.-K. Kamarainen), ilonen@lut.fi (J. Ilonen), kalviai@lut.fi (H. Kälviäinen).

0031-3203/$30.00
doi:10.1016/j.patcog.2006.01.005

fitting the restricted model to observed features. Often already simple models such as a single Gaussian distribution (normal distributed random variable) can efficiently represent patterns but a more general model, such as a finite mixture model, must be used to approximate more complex pdfs; arbitrarily complex probability density functions can be approximated using finite mixture models. The finite mixture representation is a natural choice for certain kinds of observations: observations which are produced by a randomly selected source from a set of alternative sources belonging to a same main class. This kind of task occurs when object categories are identified instead of object classes. For example, features from eye centers are partitioned into closed eye and open eye, or Caucasian and Asian eye sub-classes. The problem arises how probability densities should be approximated with finite mixture models and how the model parameters should be estimated. Equally important is to define correct practices for the use of pdfs in pattern recognition and classification tasks.

In this study Gaussian mixture model (GMM) pdfs are studied as finite mixture models. The two main considerations with the GMM are estimation of number of Gaussian components and robustness of the algorithm to tolerate singularities occurring due to a small sample size. It cannot be assumed that the user knows all necessary details, and thus, the estimation should be unsupervised and utilize existing approximation and statistical theories. Several estimation methods have been proposed in literature and the most prominent ones are experimentally evaluated in this study. The methods are extended to $\mathbb{C}^n$ since complex domain features, such as Gabor filter responses, seem to be convenient for some applications [4,7]. Correct classification practices are analyzed and defined based on problem characteristics: (i) classifying an unknown observation into one of predefined classes, (ii) finding best candidates from a set of observations, (iii) deciding class association to a single known class when other classes are unknown or their samples are insufficient, and (iv) concluding what useful statistical information should be provided to the next processing level. Finally, by providing implementations [8] for the described methods, we aim to encourage good practices when using GMM pdfs for representation and discrimination of patterns.

2. Gaussian mixture probability density function

Finite mixture models and their typical parameter estimation methods can approximate a wide variety of pdfs and are thus attractive solutions for cases where simple function forms, such as a single normal distribution, fail. However, from a practical point of view it is often sound to form the mixture using one predefined distribution type, a basic distribution. Generally the basic distribution function can be of any type but the multivariate normal distribution, the Gaussian distribution, is undoubtedly one of the most well-known and useful distributions in statistics, playing a predominant role in many areas [9]. For example, in multivariate analysis most of the existing inference procedures have been developed under the assumption of normality and in linear model problems the error vector is often assumed to be normally distributed. The multivariate normal distribution also appears in multiple comparisons, in studies of the dependence of random variables, and in many other related fields. If no prior knowledge of a pdf of a phenomenon exists, only a general model can be used and the Gaussian distribution is a good candidate. For a more detailed discussion on the theory, properties and analytical results of multivariate normal distributions we refer to Ref. [9].

2.1. Multivariate normal distribution

A non-singular multivariate normal distribution of a D dimensional random variable $\boldsymbol{X} \mapsto \boldsymbol{x}$ can be defined as

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right], \quad (1)$$

where $\boldsymbol{\mu}$ is the mean vector and Σ the covariance matrix of the normally distributed random variable $\boldsymbol{X}$. Multivariate Gaussian pdfs belong to the class of elliptically contoured distributions, and thus, for example, equiprobability surfaces of the Gaussian are $\boldsymbol{\mu}$-centered hyperellipsoids [9].

The Gaussian distribution in Eq. (1) can be used to describe a pdf of a real valued random vector ($\boldsymbol{x} \in \mathbb{R}^D$). A similar form can be derived for complex random vectors ($\boldsymbol{x} \in \mathbb{C}^D$) as [10]

$$\mathcal{N}^{\mathbb{C}}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{\pi^D|\Sigma|} \exp[-(\boldsymbol{x}-\boldsymbol{\mu})^*\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})], \quad (2)$$

where * denotes the adjoint matrix (transpose and complex conjugate).

2.2. Finite mixture model

Despite the fact that multivariate Gaussian pdfs have been successfully used to represent features and discriminate between different classes in many practical problems (e.g., Refs. [11,12]), the assumption of single component leads to strict requirements for characteristics of the phenomenon: a single basic class which smoothly varies around the class mean. The most significant problem is not typically the smooth behavior but the assumption of unimodality. For multimodally distributed features the unimodality assumption may cause an intolerable error in the estimated pdf and consequently in the discrimination between classes. Errors are not only in the limited representation power but also in completely wrong interpretations (e.g. a class represented by two Gaussian distributions and another class between them). In object recognition this can be the case for such a simple thing as a human eye, which is actually an object