PERGAMON

Pattern Recognition 34 (2001) 2015–2027

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Cluster analysis by adaptive rank-order filters

E.H. Sbai*

Université Moulay Ismaïl, Ecole Supérieure de Technologie, Route d'Agouray, BP 3103, Meknés 50006, Morocco

Received 11 March 1999; received in revised form 18 February 2000; accepted 7 April 2000

Abstract

An adaptive filter is proposed for detecting the modes of underlying probability density function of the data. The adaptive procedure is based on the selection of an appropriate rank order according to the local measurements of the entropy of the density function. The approach requires no a priori information about the structure of the data set but it is governed by the sampling parameter. Experiments demonstrate the usefulness of the filter. © 2001 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.

Keywords: Cluster analysis; Mode; Probability density function (pdf); Rank-order; Entropy

1. Introduction

Clustering of multidimensional data is an important subject in pattern recognition, particularly when only unlabeled data samples are available. Its aim is to regroup samples represented as points in multidimensional space, into homogenous clusters according to their similarities. A cluster can be loosely defined as a set of samples whose density is larger than the density of the surrounding volume. The basic techniques for clustering the data sets can be classified into two types. The first categories are called supervised, unlike the second categories called nonsupervised.

Numerous articles in the pattern analysis literature have referred to the problems of supervised and unsupervised learning [1–3]. In the context of unsupervised clustering, this paper addresses the problem of seeking the modes of the probability density function underlying the distribution of the data set. Data are represented as points in a multidimensional space with the assumption that they are drawn from a multimodal probability density function. In this context, each mode corresponds to one cluster.

When the modes are considered as local extrema of the density function, they are generally detected by means of hill climbing procedures which are known to be very sensitive to noise [4]. In the last few years, however, an alternative consisted in analysing the convexity of the density function and characterizing the modes by the convexity of this function [5]. Although this approach constitutes an interesting alternative for clustering data, it remains sensitive to local irregularities in the distribution, especially for small data sets [6].

Recently, a third approach has been proposed in which the mode detection problem is stated in terms of set theory by considering the concepts of mathematical morphology. This technique has been used to enhance cluster separability by eliminating the insignificant irregularities of their boundaries [7]. Another approach that has proved useful to improve the discrimination between the modes is the convexity-dependent morphological transformations. This procedure adapts locally the multivalue erosions and dilations to the convexity properties of the probability density function [8].

This paper investigates the application of the concepts of rank-order filters to enhance the modes and to enlarge the valleys of the multivariate probability density function when the overlapping of classes is high, especially for small data sets.

In the absence of any information concerning the underlying distribution, except, of course, the one extracted from the input patterns, nonparametric techniques of analysis are adopted [3,9–13]. In this paper, the Parzen method is used to estimate the probability density

* Tel.: + 212-5-53-85-62; fax: + 212-5-53-64-54.
E-mail address: esbai@caramail.com (E.H. Sbai).

0031-3203/01/$20.00 © 2001 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.
PII: S0031-3203(00)00130-8

function [3,9,14]. In practice, the implementation of this procedure must be undertaken with care to avoid computational burden usually associated with it. A fast algorithm is used to determine the uniform kernel estimate at a low computational cost [15].

The remainder of the paper is organized as follows. In Section 2, a brief review of a convexity-dependent morphological procedure is given and the definition of the rank-order filter is summarized. The adaptive rank-order filter is introduced in Section 3, which is the major contribution of this paper. In Section 4, numerical examples applied to artificially generated data sets and to real data are provided. The paper is concluded in Section 5.

2. Convexity-dependent morphological transformations and rank-order filters

2.1. Convexity-dependent morphological transformations

The procedure proposed by Zhang and Postaire [8] corresponds to a dilation of the underlying probability density function when it is concave and to an erosion when it is convex. This procedure makes use of a test which determines the convexity of the probability density function from the available observations.

The convexity of a density function $f(x)$ at point x_0 can be determined by the analysis of the variations of the mean value of that function computed within a family of domains expanding around the point x_0. It has been shown that this mean value is a monotonic decreasing (resp. increasing) function of the size of the domain when the expanding domain stands in a region where $f(x)$ is concave (resp. convex) [5].

Iterations of the convexity-dependent morphological operations tend to increase the amplitude of the density function in the modal regions and to decrease it in the valleys. This filtering strategy is followed by an opening operation which smoothes the contours of the modal regions and eliminates small capes and isolated peaks of the filtered input function. Unfortunately, this convexity-dependent morphological procedure followed by the opening operator gives a function with null values everywhere in the multidimensional sampling space, especially for small data sets.

In Section 3, we introduce an adaptive rank-order filter in order to overcome the limitations of the mode detection procedure based on the test of convexity and morphological operators. The approach proposed is a new application of this type of filter in the area of pattern classification.

2.2. Rank-order filters

Rank order topics reviewed here are only those necessary to establish the basic ideas to be developed in the proposed adaptive rank-order filter. The definitions and properties are rigorously given in the literature [16,17]. Rank-order filters are a class of nonlinear filters known for their robust smoothing properties in signal and image processing [18]. The rank-order filter can be easily defined on any discrete signal. Let $x_1, x_2, \ldots, x_k$, k random variables scanned through a window, or probe, of a given size $k = 3^N$(N is the dimensionality of a data space). If these variables are arranged in ascending order of magnitude as $x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(k)}$, then the rth random variable $x_{(r)}$ is the rth-order statistic ($r = 1, 2, \ldots, k$). The output value that is to be placed at the window reference origin is the rth element in the list for a rank r filter. The most important order statistics are the maximum $x_{(k)}$, the minimum $x_{(1)}$ and the median $x_{(k+1/2)}$.

Let $f(x)$, $x \in Z^N$ be a N-dimensional sampled function and H_3 be a window, which is defined as a finite subset with $|H_3| = k = 3^N$, where $|\cdot|$ denotes set cardinality. The rth order statistic (OS) of a function $f(x)$ by H_3 is the function

$$[(OS)^r(f : H_3)](x) = r\text{th}(OS)\{f(y) : y \in H_{3_x}\},$$

where $x \in Z^N$, $1 \leqslant r \leqslant k$ and $H_{3_x} = \{x + a : a \in H_3\}$ denotes the set H_3 shifted at location x. The rth rank-order filter for functions by H_3 is a filter whose output is the rth rank-order of the incoming function by H_3. The functions considered in this paper are probability density functions. In the following, we shall deal only with rank-order filters of discrete sets because the pdf is represented as a discrete set of estimated values of this function.

In the following section, an adaptive rank-order filter is developed within the general framework of rank-order filters, which are filters constrained to output an order statistic from the input samples. The adaptive filter is used to enhance the modes of the underlying probability density function and to deepen the valleys between them.

3. Design of the adaptive rank-order filter

In this section, we first show that the available data set is sampled and quantized to obtain an acceptable discrete representation of the probability density function.

Let us consider a set of Q N-dimensional observations $\{X_1, X_2, \ldots, X_q, \ldots, X_Q\}$ of a random variable X with a probability density function $f(X)$ such that

$$X_q = [X_{q,1}, X_{q,2}, \ldots, X_{q,n}, \ldots, X_{q,N}]^{\mathrm{T}}.$$

In order to extend the theory of rank-order filters to cluster analysis, this set of available observations must be represented as a mathematical discrete set in a Euclidean space. The procedure proposed here allows mapping of these observations onto the discrete space Z^N, where Z is

Download English Version:

https://daneshyari.com/en/article/533158

Download Persian Version:

https://daneshyari.com/article/533158

Daneshyari.com