Pattern Recognition 45 (2012) 3061–3071

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Statistical modeling of dissimilarity increments for *d*-dimensional data: Application in partitional clustering

Helena Aidos *, Ana Fred

Instituto de Telecomunicações, Instituto Superior Técnico, Lisbon, Portugal

ARTICLE INFO

Available online 23 December 2011

Keywords:
Dissimilarity increments
Partitional clustering
Likelihood-ratio test
Minimum description length
Gaussian mixture decomposition

ABSTRACT

This paper addresses the use of high order dissimilarity models in data mining problems. We explore dissimilarities between triplets of nearest neighbors, called *dissimilarity increments* (DIs). We derive a statistical model of DIs for *d*-dimensional data (*d*-DID) assuming that the objects follow a multivariate Gaussian distribution. Empirical evidence shows that the *d*-DID is well approximated by the particular case $d=2$. We propose the application of this model in clustering, with a partitional algorithm that uses a merge strategy on Gaussian components. Experimental results, in synthetic and real datasets, show that clustering algorithms using DID usually outperform well known clustering algorithms.

1. Introduction

Clustering has been applied in several areas like machine learning, pattern recognition, web mining, image segmentation, genetics, and biology [1–3]. The main goal of clustering is to arrange data objects in groups (clusters), such that objects belonging to the same cluster are similar. It is a form of unsupervised learning, since no information about the groups to which the objects belong is known *a priori*. Two major clustering strategies have been adopted in published methods: partitional and hierarchical [4,1,5]. Hierarchical clustering techniques group objects with a sequence of nested partitions, either from singleton clusters to a cluster including all data (agglomerative strategy) or in the opposite way (divisive strategy), while partitional clustering techniques divide the data into clusters without the hierarchical structure. For an overview of clustering techniques see [1,4–6].

Partitional methods put each data point into exactly one cluster. Often, the user must set the number of clusters, k, beforehand, and k is usually small. The choice of k can be considered itself a model selection problem [7] which is often non-trivial, especially for real-world datasets. One important class of partitional methods is the one of prototype-based methods, such as k-means [6] (with an associated minimum squared error criterion; it is the simplest and most widespread clustering algorithm), iterative self-organizing data analysis technique (ISODATA) [8], k-medoids [3] and squared-error clustering [9], which can work very well for compact and hyperspherical clusters. Another class of partitional methods is the one of parametric density approaches, including methods that estimate probability density functions from data, such as Gaussian mixture decomposition algorithms [10–12].

Hierarchical methods produce a set of nested partitions in a hierarchical structure according to a proximity matrix; this structure is graphically represented by a dendrogram [4]. Agglomerative methods start by considering each data point as one cluster, and each partition is obtained from the previous one by merging two clusters into a single cluster. Methods in this class include single-link, complete-link, average-link, median-link, centroid-link, weighted-link, Ward link [4], and more recent hierarchical algorithms for handling large-scale datasets such as CURE [13], ROCK [14], Chameleon [15] and BIRCH [16]. Divisive methods work in the opposite way: one starts with a single cluster with all the objects and a divisive procedure is applied repeatedly until all clusters are singletons. This class of methods is not very used in practice due to its computational cost: for a cluster with N objects, there are $2^{N-1}-1$ possible divisions [1]. A drawback of most classical hierarchical techniques is the failure to identify clusters with arbitrary shapes and sizes, and the tendency to form spherical structures in the data. Most of the hierarchical methods are inspired in graph theory, such as single-link and complete-link. However, graph theory can also be used in a different kind of clustering algorithms: the clusters can be described in terms of weighted graphs. CLICK [17] is an example of such methods.

Most of the clustering techniques require, implicitly or explicitly, a similarity measure between patterns, the choice of which is difficult to make if one has no prior knowledge about cluster shapes or structure. Most clustering algorithms use pairwise distances between patterns, the most typical one being the Euclidean distance. However, many other measures can be used, such as the Mahalanobis distance [1,5]. Recently, a new third

* Corresponding author. Tel.: +351 218418164.
E-mail addresses: haidos@lx.it.pt (H. Aidos), afred@lx.it.pt (A. Fred).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.12.009

order dissimilarity measure has been proposed [18], the *dissimilarity increments* (DIs), which are computed over triplets of nearest neighbor patterns. The fact that this measure uses three data points at a time gives more information about the patterns lying in the same cluster, since a smooth evolution of the DIs should occur if the patterns are in the same cluster, and high values should occur for patterns lying in different clusters [18].

Based on this new dissimilarity measure, a hierarchical clustering method has been proposed in [18]. The statistical model proposed for the DIs in a cluster, based on visual inspection, was the exponential distribution, with parameter equal to the inverse of the mean of the increments. In this paper we theoretically derive the DIs distribution (DID) under some approximations, and empirically show that this new distribution is a better approximation to the empirical distribution of the DIs than the exponential one.

The novel DID is derived under the hypothesis of local Gaussian generative models for the data in $\mathbb{R}^d$, and is called d-DID. We particularize the model for $d=2$, hereafter referred as 2-DID; using two statistical measures, we empirically show that 2-DID is a good approximation to d-DID, and that both are better approximations of the true DID than the exponential distribution. We then construct a partitional clustering algorithm consisting of a merge strategy, which iteratively accepts or rejects the merging of two clusters based on this new distribution. In [19] we proposed a likelihood-ratio test as the merge criterion, which merges pairs of clusters with a p-value less than a given significance level α. In this paper we propose a new parameter-free merge criterion based on the Minimum Description Length principle.

This paper is structured as follows: Section 2 explains the derivation of the DID for d-dimensional data (d-DID), and we write this distribution as a function of a single parameter: the expected value of the DIs. In Section 3 we present the particular case of $d=2$, and in Section 4 we show empirical evidence that this new distribution is a better approximation to the empirical distribution than the one proposed in [18]. In Section 5 we show how to use this DID in a clustering algorithm, proposing two merge criteria: likelihood-ratio test (LRT), presented in [19], and minimum description length (MDL). We present, in Section 6, the performance of the proposed algorithm on six synthetic datasets with different characteristics (Gaussian and non-Gaussian clusters, arbitrary shape clusters and densities) and on eight real-world datasets from the UCI Machine Learning Repository and 20-Newsgroups. These results are compared with a Gaussian mixture decomposition (GMD), the hierarchical clustering algorithm proposed in [18] and with some traditional clustering algorithms (single-link, average-link, complete-link, Ward-link and k-means), when the true number of clusters is known. We also present a study of the proposed method when the number of clusters is not known *a priori*. Discussion and conclusions are in Sections 7 and 8, respectively. In the Appendix, we detail the derivation of the DID.

2. Dissimilarity increments distribution for d-dimensional data (d-DID)

Consider a set of patterns X. Given $\mathbf{x}_i \in X$ and some dissimilarity measure between patterns, $d(\cdot,\cdot)$, let $(\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_k)$ be a triplet of nearest neighbors, obtained as follows: $\mathbf{x}_j$ is the nearest neighbor of $\mathbf{x}_i$, and $\mathbf{x}_k$ is the nearest neighbor of $\mathbf{x}_j$ different from $\mathbf{x}_i$. The *dissimilarity increment* (DI) [18] between these patterns is defined as

$$d_{inc}(\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_k) = \left|d(\mathbf{x}_i,\mathbf{x}_j) - d(\mathbf{x}_j,\mathbf{x}_k)\right|. \tag{1}$$

In the following subsections we will derive the probability density function (PDF) for the DIs, using the Euclidean distance as the dissimilarity measure.

2.1. Derivation of the DID model

Assume that X is a d-dimensional set of patterns (henceforth called a *cluster*), and that its elements are independent and identically distributed according to a multivariate Gaussian distribution, $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu},\Sigma)$. With no loss of generality, we assume that $\boldsymbol{\mu}=0$ and that Σ is diagonal (this only involves translation and rotation of the data, which does not affect Euclidean distances). If $\mathbf{x}$ denotes a sample from this Gaussian, we define the sphered data $\mathbf{x}^*$ as having its i-th entry given by $x_i^* \equiv x_i/\Sigma_{ii}$ (this transformation is known as "whitening" or "sphering"); x_i^* thus follows the standard normal distribution, $\mathcal{N}(0,1)$. The difference between samples from two univariate standard normal distributions follows a normal distribution with covariance 2. It can be shown that the squared Euclidean distance, $(D^*)^2 = \sum_{i=1}^{d}(z_i^*)^2$, where $z_i^* \equiv (x_i^* - y_i^*)/\sqrt{2} \sim \mathcal{N}(0,1)$, follows a chi-square distribution with d degrees of freedom [20]. Thus, the PDF for $(D^*)^2$ is given by

$$p_{(D^*)^2}(x) = \frac{2^{-d/2}}{\Gamma(d/2)} x^{d/2-1} \exp\left(-\frac{x}{2}\right), \quad x \in [0,+\infty[, \tag{2}$$

where $\Gamma(\cdot)$ denotes the gamma function.

Furthermore, after the sphering, the transformed data has circular symmetry in $\mathbb{R}^d$. We define angular coordinates in a $(d-1)$-sphere, with $\theta_i \in [0,\pi[, i=1,\ldots,d-2$ and $\theta_{d-1} \in [0,2\pi[$. Define $\mathbf{D} \equiv \mathbf{x}-\mathbf{y} \equiv (b_1,b_2,\ldots,b_d)$, where b_i can be expressed in terms of polar coordinates as

$$b_1 = \sqrt{2\Sigma_{11}} D^* \cos\theta_1,$$

$$b_i = \sqrt{2\Sigma_{ii}}\, D^* \left[\prod_{k=1}^{i-1} \sin\theta_k\right] \cos\theta_i, \quad i=2,\ldots,d-1,$$

$$b_d = \sqrt{2\Sigma_{dd}}\, D^* \left[\prod_{k=1}^{d-1} \sin\theta_k\right].$$

The squared Euclidean distance in the original space is

$$\begin{aligned} D^2 &= 2\left[\Sigma_{11}\cos^2\theta_1 + \sum_{i=2}^{d-1}\Sigma_{ii}\left(\prod_{k=1}^{i-1}\sin^2\theta_k\right)\cos^2\theta_i \right. \\ &\quad \left. + \Sigma_{dd}\left(\prod_{k=1}^{d-1}\sin^2\theta_k\right)\right](D^*)^2 \\ &\equiv 2A(\Theta)(D^*)^2, \end{aligned} \tag{3}$$

where $A(\Theta)$, with $\Theta = (\theta_1,\theta_2,\ldots,\theta_{d-1})$, is called the *expansion factor*. Naturally, this expansion factor depends on the angle vector Θ. In practice, it is hard to properly deal with this dependence; so, we will use the approximation that the expansion factor is constant and equal to the expected value of the true expansion factor over all angles Θ. This expected value is given by

$$\mathbb{E}[A(\Theta)] = \frac{\pi^{-d/2+1}}{2\Gamma\left(1+\frac{d}{2}\right)}\eta, \tag{4}$$

where $\eta \equiv \mathrm{tr}(\Sigma)$ (see Appendix A for the derivation).

Under this approximation, the transformation equation (3) from the normalized space to the original space is given by

$$D^2 = \frac{\pi^{-d/2+1}}{\Gamma(1+d/2)}\eta(D^*)^2. \tag{5}$$

From (2) and (5) one can obtain the PDF of D^2, and from there one can obtain the PDF of $D = d(\mathbf{x},\mathbf{y})$ as

$$p_D(y) = 2G_d(\eta)y^{d-1}\exp\left(-C_d(\eta)y^2\right), \quad y \in [0,+\infty[, \tag{6}$$

where we define $G_d(\eta) \equiv d^{d/2}\Gamma(d/2)^{d/2-1}2^{-d}\eta^{-d/2}\pi^{d/2(d/2-1)}$ and $C_d(\eta) \equiv d\Gamma(d/2)(4\eta)^{-1}\pi^{d/2-1}$.

The DI is defined as the absolute value of the difference of two Euclidean distances. We have just derived the PDF of the