Pattern Recognition 47 (2014) 2031–2041

Contents lists available at ScienceDirect

Pattern Recognition

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

Asymmetric clustering using the alpha–beta divergence

Dominik Olszewski [a,*], Branko Šter [b]

[a] Faculty of Electrical Engineering, Warsaw University of Technology, Koszykowa 75 Street, 00-662 Warsaw, Poland
[b] Faculty of Computer and Information Science, University of Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 10 October 2012
Received in revised form
16 November 2013
Accepted 20 November 2013
Available online 1 December 2013

Keywords:
Clustering
Asymmetry
Dissimilarity
Alpha–Beta divergence

ABSTRACT

We propose the use of an asymmetric dissimilarity measure in centroid-based clustering. The dissimilarity employed is the Alpha–Beta divergence (AB-divergence), which can be asymmetrized using its parameters. We compute the degree of asymmetry of the AB-divergence on the basis of the within-cluster variances. In this way, the proposed approach is able to flexibly model even clusters with significantly different variances. Consequently, this method overcomes one of the major drawbacks of the standard symmetric centroid-based clustering.

1. Introduction

In everyday life, people are used to consider the dissimilarity between two entities as a symmetric relation. In many cases, it is indeed symmetric. If it also satisfies the triangle inequality, it is called metric or simply distance. Besides measuring a physical distance, it can also serve as a measure of dissimilarity.

However, a dissimilarity does not need to be symmetric. One of the first to come upon this idea was Amos Tversky, who questioned the geometric representation of similarity [1]. He argued that the notion of similarity had been dominated by geometric models, which represent objects as points in some coordinate space and that dissimilarities between objects simply correspond to the metric distances between the points. He argues that a similarity statement, such as "a is like b", is directional. It has a subject and a referent, and is not equivalent to the statement "b is like a". His well-known example states that "North Korea is more similar to China than China to North Korea", since China is a larger and a more general entity. Or, we say "the son resembles the father" rather that "the father resembles the son", since the father is the more prominent entity. His claims were validated in his numerous psychological experiments [2], and his idea was undoubtedly an inspiration for many later works concerning the asymmetric dissimilarities and the general problem of asymmetry in data analysis.

A direct relation to Tversky's finding can be found in the work of Nosofsky [3]. In contrast to the asymmetric form of similarity, Nosofsky proposes to use a differential "bias", which will be assigned to individual objects, in contrary to similarity, which is always determined between two objects. Employing this kind of "bias" can be considered as equivalent to applying an asymmetric similarity. And, according to Nosofsky, the "bias-based" models may be even superior over the standard symmetric-similarity-based models. The author even expresses his surprise by the fact that symmetric similarity has gained the dominant position in the world of data analysis.

A further continuation of the mentioned ideas appears in the paper [4] by Johannesson, where the asymmetric relationship between two objects is expressed using the traditional symmetric similarity and the quotient of "prominences" between those objects. The author compares his results with those obtained with usage of the "bias-based" approach presented by Nosofsky.

Another similar concept closely related to the idea of Tversky appears in the work of Martín-Merino and Muñoz [5], where the asymmetric version of the Self-Organizing Map was proposed. The authors notice the same asymmetric directional relationships between objects of different levels of generality (or prominence). Their example from the field of textual data analysis concerns the dissimilarity between the two words – "mathematics" and "Bayes". The former is a more general entity, which makes the relationship between those words strongly asymmetric. Symmetric dissimilarities produce large values for most pairs of objects, and consequently, they do not reflect properly the associations between objects of different levels of generality. As it is stated in [5], asymmetry can be interpreted as a particular type of hierarchy. In [6], Martín-Merino and Muñoz also find the cause

* Corresponding author. Tel.: +48 22 234 7618; Tel./fax: +48 22 625 6278.
E-mail addresses: dominik.olszewski@ee.pw.edu.pl (D. Olszewski), branko.ster@fri.uni-lj.si (B. Šter).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.11.019

of asymmetric nature of data in hierarchical relationships between objects. Diego et al. [7] combine several similarity matrices into one kernel and train a Support Vector Machine.

A continuation of the idea of hierarchical-caused asymmetry can be found in [8], where the asymmetric version of the *k*-means clustering algorithm was introduced. The author utilized a similar assertion justifying the usage of asymmetric dissimilarities. Also in [9], where the improved version of the asymmetric *k*-means algorithm using the asymmetric coefficients was proposed, the asymmetric dissimilarity was employed as preferable over the standard symmetric one. In [9], the asymmetric coefficients were used in order to determine the degree of asymmetry of dissimilarity. This assured that the improved asymmetric *k*-means algorithm properly adjusts to the properties of the analyzed dataset.

Another direction of research arose focused on the issue of aggregation of single dissimilarities resulting in obtaining a novel complex dissimilarity measure. In [10], the aggregate asymmetric dissimilarity is computed as a weighted sum of a fixed number of input dissimilarities, while in [11], an appropriate asymmetric distance aggregation function is formulated and employed.

In general, the problem of asymmetry in data analysis was also studied by Okada and Imaizumi [12–15]. Their work is focused on using the dominance point governing asymmetry in the proximity relationships among objects, represented as points in a multidimensional Euclidean space. They claim that ignoring or neglecting the asymmetry in proximity analysis discards potentially valuable information. On the other hand, Zielman and Heiser in [16] consider the models for asymmetric proximities as a combination of a symmetric similarity component and an asymmetric dominance component.

1.1. Our proposal

In this paper, we propose an asymmetric centroid-based clustering approach using the asymmetric dissimilarity. As the asymmetric dissimilarity, we have used the Alpha–Beta divergence (AB-divergence), introduced recently in [17]. The formula of this dissimilarity measure involves parameters, which can be used to tune the degree of asymmetry of the dissimilarity. Therefore, this quantity was particularly useful in our research. The values of the appropriate parameters are computed based on the variances of the clusters in the analyzed dataset. In this way, the geometric or areal sizes of the clusters in the feature space are taken into account, thereby overcoming one of the well-known drawbacks of the traditional centroid-based clustering. The notion of the size of a cluster refers here to the area occupied by the objects of that cluster in the feature space. In other words, the size of a cluster in this work is related to the variance of objects in that cluster, i.e., the within-cluster variance. As it is explained in Section 4, the proposed approach assures that the dissimilarity in the clustering process is determined more accurately than in case of standard symmetric quantities (for example the Euclidean distance), and consequently, a higher clustering performance is obtained.

The results of the experimental study conducted on real and simulated data of high and low dimensionality confirm the effectiveness of the proposed approach, which in most cases outperforms four other investigated clustering methods.

1.2. Remainder of the paper

The rest of this paper is organized as follows: in Section 2, the centroid-based clustering approach is described; Section 3 presents the AB-divergence and its most important properties; in Section 4, the usage of the AB-divergence in the centroid-based clustering, which constitutes the main proposal of the paper, is explained; Section 5 reports the results of the experimental study on three different datasets together with the discussion of the results; while Section 6 summarizes the whole paper, and provides concluding remarks.

2. Centroid-based clustering

Data clustering process aims to form clusters of possibly most similar objects in a given analyzed dataset. An object represented as a vector of d features can be interpreted as a point in the d-dimensional space. A centroid-based clustering algorithm (also known as k-centroids clustering algorithm [18–21]) is a statistical data analysis tool used in order to form an arbitrary settled number of clusters in the analyzed dataset. It can be formulated as follows: given n points in a d-dimensional space and the number of desired clusters k, the algorithm searches for a set of k clusters so as to minimize the sum of squared distances or dissimilarities between each point and its cluster centroid. The cluster centroid is a point being possibly the best representation of the whole cluster.

The k-centroids clustering algorithm consists of two alternating steps:

Step 1. Forming of the clusters: the algorithm iterates over the entire dataset and allocates each object to the cluster represented by the centroid – nearest to this object. The nearest centroid is determined with the use of a selected dissimilarity measure. Hence, for each object in the analyzed dataset, the following minimal squared dissimilarity has to be found:

$$\min_j D^2(x_i \| c_j), \tag{1}$$

where $D(\cdot\|\cdot)$ is a selected dissimilarity measure, x_i, $i=1,\ldots,n_j$, is an object in the jth cluster, c_j, $j=1,\ldots,k$, is the centroid of the jth cluster, and n_j, $j=1,\ldots,k$, is the number of objects in the jth cluster.

Step 2. Finding centroids of the clusters: for each cluster, a centroid is determined on the basis of objects belonging to this cluster. The algorithm calculates centroids of the clusters so as to minimize a formal objective function, the error distortion:

$$e(X_j) = \sum_{i=1}^{n_j} D^2(x_i \| c_j), \tag{2}$$

where X_j, $j=1,\ldots,k$, is the jth cluster, x_i, $i=1,\ldots,n_j$, is the object in the jth cluster, c_j, $j=1,\ldots,k$, is the centroid of the jth cluster, n_j, $j=1,\ldots,k$, is the number of objects in the jth cluster, k is the number of clusters, and $D(\cdot\|\cdot)$ is the selected dissimilarity measure.

Both these steps must be carried out with the same dissimilarity measure, in order to guarantee the monotone property of the k-centroids algorithm.

Steps 1 and 2 have to be repeated until the termination condition is met. The termination condition might be either reaching convergence of the iterative application of the objective function, or reaching the pre-defined number of cycles.

After each cycle (Steps 1 and 2), the value of the error function – expressing the scattering of objects in the entire dataset – needs to be computed for the entire analyzed dataset, in order to track the convergence of the whole clustering process:

$$e(X) = \sum_{j=1}^{k} \sum_{i=1}^{n_j} D^2(x_i \| c_j), \tag{3}$$