Pattern Recognition 48 (2015) 2812–2830

Contents lists available at ScienceDirect

Pattern Recognition

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

Kernel methods for point symmetry-based clustering

Guillaume Cleuziou [a,b,*], Jose G. Moreno [b]

[a] Université d'Orléans, INSA Centre Val de Loire, LIFO Bat. 3IA, LIFO EA 4022, Rue Léonard de Vinci, B.P. 6759, FR-45067 Orléans Cedex 2, France
[b] Université de Caen Basse-Normandie, GREYC UMR 6072, FR-14032 Caen, France

ARTICLE INFO

Article history:
Received 12 August 2014
Received in revised form
15 February 2015
Accepted 16 March 2015
Available online 24 March 2015

Keywords:
Pattern recognition
Clustering
Point symmetry-based distance measure
Kernel function
K-means

ABSTRACT

This paper deals with the point symmetry-based clustering task that consists in retrieving – from a data set – clusters having a point symmetric shape. Prototype-based algorithms are considered and a non-trivial generalization to kernel methods is proposed, thanks to the geometric properties satisfied by the point symmetry distances proposed until now. The proposed kernelized framework offers new opportunities to deal with non-Euclidean symmetries and to reconsider any intractable examples by means of implicit feature spaces.

A deep experimental study is proposed that brings out, on artificial data sets, the capabilities and the limits of the current point symmetry-based clustering methods. It reveals that kernel methods are quite capable of stretching the current limits for the considered task and encourages new research on the kernel selection issue in order to design a fully unsupervised symmetric pattern recognition process.

1. Introduction

Clustering is a well-known task in data mining and pattern recognition that consists in organizing a set of objects into groups (or clusters) in such a way that similar objects belong to the same cluster and dissimilar objects belong to different clusters. Originally, clustering was used in the data analysis process to build summaries (or typologies) from data sets, usually stored in single tables (*objects* × *features*). It gave rise to intensive research during the last decades by studying numerous strategies to define and/or to reach good clustering solutions: model-based clustering, partitioning methods, hierarchical algorithms, density-based or graph-based approaches to name but a few [1–3]. Then, with the development of the digital technologies, the application domains and the diversity of the data types have increased, making data clustering a challenging task. At the same time, the requirements on the emerging solutions have grown in such a way that the metrics, the models and/or the algorithms usually employed have to be redesigned in order to take into account the user's or domain's expectations. For example, when clustering is used to control the topology of a sensor network by organizing the sensor nodes, the efficiency of the network needs to build clusters with either balanced [4] or unbalanced [5] sizes; organizing a set of genes according to their metabolic functions naturally requires considering overlapping clusters rather than crisp-partitionings [6–8]; if the user has predefined requirements on whether some objects must or must-not belong to a same cluster, his partial knowledge can be used to drive either the metric [9], the representation space [10] or the clustering process itself [11] with a semi-supervised learning strategy; the shape of the expected clusters is also subject to studies that aim to retrieve not only spherical clusters (as with the famous *k*-means algorithm) [12] but also ellipsoidal [13,14] and non-convex cluster shapes [15].

This paper focuses on the last issue that aims at retrieving clusters with particular shapes – namely symmetrical patterns. Observing that many physical things that surround us have exact or approximative geometrical properties, symmetrical pattern recognition is of high interest for example in computer vision. This task has been tackled using clustering methodologies through the starting work in 2001 from Su and Chou [16] who proposed a new (nonmetric) "point symmetry" distance and use it with a modified *k*-means algorithm in order to retrieve clusters that present a symmetrical structure with respect to the cluster center (circles, rings, bands, stars, etc.). Improvements on both the distance measure and the clustering process have then been proposed in order to

1. deal with situations where clusters themselves are placed in a (point) symmetric manner [17,18],
2. better explore the solution space that is claimed to be inefficiently scanned with the original reallocation algorithm [18,19].

On the whole, the previous approaches succeed in the recognition of symmetrical clusters, as illustrated with the experiments led on both artificial data sets and real images [16–19]. However, the main issue in any of the current symmetry based clustering methods is the fact that they all consider the Euclidean distance as basic information

* Corresponding author at: Université d'Orléans, INSA Centre Val de Loire, LIFO Bat. 3IA, LIFO EA 4022, Rue Léonard de Vinci, B.P. 6759, FR-45067 Orléans Cedex 2, France. Tel.: +33 2 38 49 25 91; fax: +33 2 38 41 71 37.
E-mail address: guillaume.cleuziou@univ-orleans.fr (G. Cleuziou).

http://dx.doi.org/10.1016/j.patcog.2015.03.013
0031-3203/

in the computation of the proposed point symmetry distances, thus restricting the approaches to make objects that have natural structuring into clusters with Euclidean symmetrical shapes. In application domains where the Euclidean distance is not well adapted to quantify the closeness between objects, the previous approaches cannot be used, whereas symmetrical clusters can appear by using a suitable distance or similarity measure.

Kernelization is a powerful mathematical tool that enables us to perform implicit projections on the data set thus making a clustering method like *k*-means able to retrieve clusters with nonlinear separating hypersurfaces [20,21]. In the same time this process generalizes a clustering method designed for a specific proximity measure (typically the Euclidean distance) to any similarity measure between objects that can be formalized as a positive semi-definite matrix [22]. The aim of the present paper is to extend actual point symmetry distances to high-dimensional spaces using kernel functions in order to generalize the current point symmetry based clustering models.

We first recall in Section 2 the definitions and the limitations of the three main point symmetry distances proposed in [16,17] and [19]; they are all based on the Euclidean distance between a mirror image point to compute in the original Euclidean space and it(s) nearest neighbor(s). Such mirror image points make non-trivial the kernelization of the point symmetry distances. We propose in Section 3 a geometrical reasoning that enables us to reformulate each of the three point symmetry distances as expressions using only scalar products between initial objects thus leading to a kernelized clustering process described in Section 4. The two following sections are devoted to experiments on simulated data sets: in Section 5 we show how the new kernelized method can effectively retrieve symmetrical clusters for objects originally compared with a non-Euclidean distance measure; in Section 6 we take benefit from the projection capabilities offered by the new kernel-based clustering method and we show that such projections can be efficiently exploited to deal with data sets for which non-kernelized approaches fail to retrieve reliable symmetric clusters, even when symmetric clusters neatly appear from the Euclidean (natural) description of the objects. Finally, Section 7 concludes the paper.

2. Point symmetry distances and clustering algorithms

In the following we consider a data set $X=\{x_1,\ldots,x_N\}$ containing N objects to organize into K symmetrical clusters $\Pi=\{\pi_1,\ldots,\pi_K\}$. When objects can be defined as vectors in a P-dimensional space $\mathbb{R}^P$, we denote as c_k the center (or centroid) of cluster π_k also defined in $\mathbb{R}^P$ as $(c_{k,1},\ldots,c_{k,P})^T$ with

$$c_{k,v}=\frac{\sum_{x_i\in\pi_k}x_{i,v}}{|\pi_k|} \tag{1}$$

To compare two points in $\mathbb{R}^P$, the Euclidean distance is commonly used as proximity measure:

$$\|x_i-x_j\|=\sqrt{\sum_{v=1}^{P}(x_{i,v}-x_{j,v})^2} \tag{2}$$

Using the previous notations, we review in the two following subsections, first the point symmetry distances proposed in [16,17] and [19] and then, the clustering algorithms used to capture symmetrical clusters.

2.1. Point symmetry distances

Su and Chou defined in [16] a first so-called "point symmetry" distance measure that quantifies whether a pattern $x_i\in X$ is a good candidate to be member of a symmetrical cluster $\pi_k\in\Pi$, represented by its centroid c_k. To be a good candidate, the object x_i must have a (almost) symmetrical object in X with respect to c_k. Such a symmetrical object, denoted $x_i^{c_k}{}^*$ in the following, is defined as the nearest object from its exact mirror point $\overline{x}_i^{c_k}$. Fig. 1 illustrates the notions of *mirror point* and *symmetrical object*: $\overline{x}_i^{c_k}$ is the (exact) mirror point of x_i with respect to the center c_k, it is defined by $\overline{x}_i^{c_k}=(2c_k-x_i)$; and the symmetrical object $x_i^{c_k}{}^*$ is the object from X that is the nearest neighbor from the mirror point:

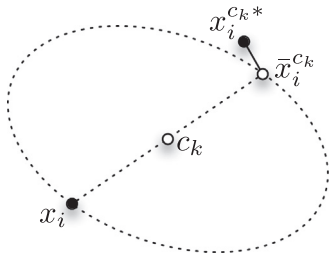

Fig. 1. Illustration of the *mirror point* $\overline{x}_i^{c_k}$ and the *symmetrical object* $x_i^{c_k}{}^*$ of x_i with respect to a cluster center c_k in the point symmetry distance computation.

$$x_i^{c_k}{}^*=\underset{x_j\in X, x_j\neq x_i}{\operatorname{argmin}}\ \|x_j-\overline{x}_i^{c_k}\| \tag{3}$$

For a better understanding of the notations and illustrations, let us mention that *bar* notations (e.g. $\overline{x}_i^{c_k}$) and *white* illustrative points in the figures (○) are used to indicate points in the $\mathbb{R}^P$ space that are not necessarily present in the data set X, whereas *unbar* notations (e.g. $x_i^{c_k}{}^*$) and *black* illustrative points in the figures (●) refers to objects from X.

Whatever the hidden symmetrical shape of the cluster π_k (an ellipse in our illustration), the distance $\|\overline{x}_i^{c_k}-x_i^{c_k}{}^*\|$ must be as small as possible to make x_i a good candidate for π_k. The point symmetry distance[1] proposed in [16] is formalized by

$$d_s(x_i,c_k)=\min_{j=1..N, j\neq i}\frac{\|(x_i-c_k)+(x_j-c_k)\|}{\|x_i-c_k\|+\|x_j-c_k\|} \tag{4}$$

Observing that the minimization of the numerator leads actually to the distance between the mirror point $\overline{x}_i^{c_k}$ and its nearest neighbor $x_i^{c_k}{}^*$ in the data set, Bandyopadhyay and Saha [19] rewrote the Su and Chou's distance in the following manner:

$$d'_s(x_i,c_k)=\frac{\|\overline{x}_i^{c_k}-x_i^{c_k}{}^*\|}{\|x_i-c_k\|+\|x_i^{c_k}{}^*-c_k\|} \tag{5}$$

In fact, contrary to what is claimed in [19], Eqs. (4) and (5) are not strictly equivalent since the minimization is not only about the numerator but concerns the whole term in such a way that an other pattern $x_j\in X$ could be a better minimizer than $x_i^{c_k}{}^*$ because of its higher distance with the cluster center c_k (part of the denominator). But this phenomenon can be considered as a side effect of the denominator that was initially introduced as a single normalization term, and the modified definition of [19] must be seen as a (probably unintentional) small improvement of the Su and Chou's distance. Thus, in the following we will use Eq. (5) to denote the point symmetry distance.

A second, and more noticeable, side effect of the normalization term in both Eqs. (4) and (5) appears when the point symmetry distance is used to compare the assignment of a pattern x_i to different clusters. Chou et al. [17] first notice that, when almost symmetrical objects exist for several clusters, the denominator favors the assignment to the farthest one. This phenomenon typically occurs when

[1] It is worth noting at this stage that this measure and any of the symmetric proximity measures that will be defined in the following are called abusively "distance", to the extent that they do not satisfy to the basic mathematical requirements for such a metric like symmetry, identity or minimality.