Applied Soft Computing 12 (2012) 1542–1554

Contents lists available at SciVerse ScienceDirect

Applied Soft Computing

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

DBCAMM: A novel density based clustering algorithm via using the Mahalanobis metric

Yan Ren[a,b,*], Xiaodong Liu[a], Wanquan Liu[c]

[a] Research Center of Information and Control, Dalian University of Technology, Dalian 116024, PR China
[b] College of Automation, Shenyang Aerospace University, Shenyang 110136, PR China
[c] Department of Computing, Curtin University, Perth, WA 6102, Australia

ARTICLE INFO

Article history:
Received 5 April 2011
Received in revised form 20 October 2011
Accepted 12 December 2011
Available online 5 January 2012

Keywords:
Clustering
Mahalanobis distance
Leaders
Followers
Image segmentation

ABSTRACT

In this paper we propose a new density based clustering algorithm via using the Mahalanobis metric. This is motivated by the current state-of-the-art density clustering algorithm DBSCAN and some fuzzy clustering algorithms. There are two novelties for the proposed algorithm: One is to adopt the Mahalanobis metric as distance measurement instead of the Euclidean distance in DBSCAN and the other is its effective merging approach for leaders and followers defined in this paper. This Mahalanobis metric is closely associated with dataset distribution. In order to overcome the unique density issue in DBSCAN, we propose an approach to merge the sub-clusters by using the local sub-cluster density information. Eventually we show how to automatically and efficiently extract not only 'traditional' clustering information, such as representative points, but also the intrinsic clustering structure. Extensive experiments on some synthetic datasets show the validity of the proposed algorithm. Further the segmentation results on some typical images by using the proposed algorithm and DBSCAN are presented in this paper and they are shown that the proposed algorithm can produce much better visual results in image segmentation.

1. Introduction

Clustering techniques have many applications, such as image segmentation, information retrieval and computer vision. Therefore, it is not surprising to see the continued popularity of data clustering in research [1]. An ideal cluster can be defined as a set of points that is compact and isolated, satisfying some hypothesis. In reality, a cluster is a subjective entity with the belief of the beholder and its significance and interpretation require domain knowledge.

Among the clustering methods, DBSCAN [2] (Density-Based Spatial Clustering of Applications with Noise) is an effective clustering algorithm for spatial database systems [3] due to its capability of recognizing clusters with arbitrary shapes. Roughly speaking, it defines a cluster to be a set with maximum number of density-connected data points, in which every core data point must have at least a minimum number of data points (MinPts) within a neighbor of given radius (Eps). DBSCAN can find arbitrarily shaped clusters with the cluster density being determined beforehand and also the cluster density is uniform for all points in a dataset. However, DBSCAN is very sensitive to the selection of the two parameters MinPts and Eps, i.e., a slightly different setting may lead to total different partitions of a dataset [4]. For such reasons, some variants of DBSCAN are proposed recently to overcome this issue [5].

Instead of clustering a dataset into different non-overlap groups, fuzzy clustering algorithms could partition a dataset into overlapping groups such that these clusters describe an underlying structure within the dataset [6]. Among these fuzzy clustering approaches, the objective function-based fuzzy clustering algorithms have been received extensive attention recently, and the intention is to partition a dataset into a specified number of clusters, no matter whether the produced clusters are meaningful or not. The number of clusters should ideally correspond to the number of substructures naturally present in the dataset. For such purpose, many methods [7] have been proposed to determine the relevant number of clusters. Furthermore, the kernel-based augmentation of these objective function-based fuzzy clustering algorithms has become popular in the communities of data mining and machine learning with an aim to solve the classification and regression problems, for example, the Fuzzy C-Means (FCM) [6] and its generalizations of Gustafson–Kessel (GK) FCM [8–10]. The performances of these clustering algorithms in the kernel-induced feature space have been evaluated based on various datasets [11]. The evaluation results show that each kernel-based clustering algorithm works better than its original algorithm for almost all the data sets used in the experiments.

Furthermore, all clustering techniques are based on a metric between objects. The most two commonly used distance measures

* Corresponding author at: Research Center of Information and Control, Dalian University of Technology, Dalian 116024, PR China.
E-mail address: renyan1108@yahoo.com.cn (Y. Ren).

1568-4946/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.asoc.2011.12.015

are Euclidean distance and Mahalanobis distance [12]. For example, the DBSCAN calculates the Euclidean distances between points in order to select the optimal clustering result, and the FCM algorithm also uses the Euclidian metric [13]. The GK algorithm extends the FCM algorithm by utilizing the Mahalanobis distance. In fact, many algorithms need to be equipped with a suitable distance metric, through which the neighboring data points can be identified. Intuitively, the Euclidean distance metric implies that each data point is equally important and independent from others, and this may not be always true in applications. In contrast, a good distance metric should identify important features and discriminate relevant and irrelevant features. Thus, selecting such a good distance metric is highly problem-specific and it may determine the success or failure of a clustering algorithm or the developed system [14–17]. The Mahalanobis distance takes into account the correlation within the two variables or attributes and it serves an ideal choice in this paper.

In this paper, we propose a novel clustering algorithm based on the Mahalanobis distance named as DBCAMM (density based clustering algorithm with Mahalanobis metric). Briefly speaking for the clustering process, DBCAMM derives the leaders in succession and the points in the leader's neighbors are viewed as followers. The distance metric for selecting neighbor points is by using the Mahalanobis distance between two points, in which the distribution of the dataset will be taken into account. Consequently all leaders will form a chain, and the former is the latter's leader, i.e., the leaders have a sequence of order. Then a hierarchical clustering approach is designed, which starts with the first leader and its followers. Pairs of the followers of the new derived leader and the existing clusters are then successively merged by using the local neighbors' information until all points are assigned into a cluster. Compared with DBSCAN, DBCAMM needs not to use the concepts of "the density-reachability" and "the density-connectivity", which depend on a unified density for whole dataset. DBCAMM needs not to define any local or global density parameters like some DBSCAN variants [18]. Especially, experimental results demonstrate that DBCAMM not only outperforms the related existing algorithms in performance but also has lower computational complexity than DBSCAN. Compared with the objective function-based fuzzy clustering algorithms, DBCAMM also possesses lower computational complexity and experimental results confirm that the quality of clustering produced by DBCAMM is much better than those produced by the existing objective function-based fuzzy clustering algorithms.

The remainder of this paper is organized as follows. Section 2 will briefly introduce the related works. In Section 3, the proposed algorithm of DBCAMM is presented. The experimental results and the empirical comparisons on synthetic examples are shown in Section 4. The results on image segmentation are reported in Section 5. Section 6 concludes this paper.

2. Related works

2.1. DBSCAN

Since the proposed DBCAMM is a density based clustering algorithm and it is closely related to DBSCAN. We first briefly introduce DBSCAN in this section in order to compare them effectively. In fact, DBSCAN [2] is designed to discover clusters and noise points for Spatial Database Systems. It defines a cluster to be a maximum set of density-connected data points, in which every core data point in a cluster must have at least a minimum number of data points (*MinPts*) within a neighbor of given radius (*Eps*). After DBSCAN clustering, all data points within one cluster can be reached from one to another by traversing a path of density-connected data points while the data points across different clusters cannot. DBSCAN can find arbitrarily shaped clusters. However, DBSCAN is very sensitive to the selection of two parameters *MinPts* and *Eps*, i.e., slightly different setting of them may lead to very different partitions of dataset [4]. In order to describe DBSCAN in more detail, we need the following concepts [2].

In DBSCAN, the distance of two points is determined by a distance metric, such as the Euclidean distance. For two points p and q in a dataset D, the distance between them is denoted by $dist(p, q)$. Usually the distance is only dependent on these two points and independent on the dataset distribution.

Definition 1. (Eps-neighborhood). The Eps-neighborhood of a point p is defined by $\{q \in D | dist(p, q) \leq Eps\}$.

Definition 2. (Core point). A core point contains at least a minimum number (MinPts) of other points within its Eps-neighborhood.

Definition 3. (Directly density-reachable). A point p is directly density-reachable from a point q if p is within the Eps-neighborhood of q, and q is a core point.

Definition 4. (Density-reachable). A point p is density-reachable from the point q with respect to Eps and MinPts if there is a chain of points $p_1, \ldots, p_n$, $p_1 = q$ and $p_n = q$ such that p_{i+1} is directly density-reachable from p_i with respect to Eps and MinPts, for $1 \leq i \leq n, p_i \in D$.

Definition 5. (Density-connected). A point p is density-connected to a point q with respect to Eps and MinPts if there is a point $o \in D$ such that both p and q are density-reachable from o with respect to Eps and MinPts.

Definition 6. (Density-based cluster). A cluster C is a non-empty subset of D satisfying the following "maximality" and "connectivity" requirements:

(1) $\forall p, q$: if $q \in C$ and p is density-reachable from q with respect to Eps and MinPts, then $p \in C$.
(2) $\forall p, q \in C$: p is density-connected to q with respect to Eps and MinPts.

Definition 7. (Border point). A point p is a border point if it is not a core point but density-reachable from another core point.

With these concepts, one can state the following DBSCAN algorithm as follows.

DBSCAN

Step 1. Select the two parameters Eps and MinPts.
Step 2. Mark all the points in the dataset as unclassified and set $t = 1$.
Step 3. Find an unclassified core-point p with parameters Eps and MinPts. Mark the point p to be classified. Start a new empty cluster C_t and assign p to this cluster.
Step 4. Find all the unclassified points in the Eps-neighborhood of p and define them seed points.
Step 5. Take a point q in the set of seed points, mark q to be classified, assign q to the cluster C_t, and remove q from the set of seed points.
Step 6. Check if q is a core-point with parameters Eps and MinPts, if so, add all the unclassified points in the Eps-neighborhood of q to the set of seed points.
Step 7. Repeat steps 5 and 6 until the set of seed points is empty.
Step 8. Set $t = t + 1$ and repeat steps 3–7 until no more core points can be found.
Step 9. Output all the clusters found so far; and mark all the points which do not belong to any cluster as noise points.

End.