Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 41 (2008) 2757–2776
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

Automatic clustering and boundary detection algorithm based on adaptive influence function

Gleb V. Nosovskiy[a], Dongquan Liu[b], Olga Sourina[b,*]

[a] *Faculty of Mechanics and Mathematics, Moscow State University, Moscow GSP-2, Russia*
[b] *School of Electrical and Electronic Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798*

Received 15 January 2007; received in revised form 17 January 2008; accepted 30 January 2008

Abstract

Clustering became a classical problem in databases, data warehouses, pattern recognition, artificial intelligence, and computer graphics. Applications in large spatial databases, point-based graphics, etc., give rise to new requirements for the clustering algorithms: automatic discovering of arbitrary shaped and/or non-homogeneous clusters, discovering of clusters located in low-dimensional hyperspace, detecting cluster boundaries. On that account, a new clustering and boundary detecting algorithm, ADACLUS, is proposed. It is based on the specially constructed adaptive influence function, and therefore, discovers clusters of arbitrary shapes and diverse densities, adequately captures clusters boundaries, and it is robust to noise. Normally ADACLUS performs clustering purely automatically without any preliminary parameter settings. But it also gives the user an optional possibility to set three parameters with clear meaning in order to adjust clustering for special applications. The algorithm was tested on various two-dimensional data sets, and it exhibited its effectiveness in discovering clusters of complex shapes and diverse densities. Linear complexity of the ADACLUS gives it an advantage over some well-known algorithms.

Keywords: Clustering algorithms; Data mining; Density-based clustering

1. Introduction

Clustering is a classical problem in databases, data warehouses, pattern recognition, and artificial intelligence. In last 10 years, clustering algorithms were significantly improved, and many new features were implemented in them. Recently, due to development of point-based graphics in computer geometry, a new application area for clustering algorithms has emerged [1–3]. It brought new requirements to clustering procedures. To fit the geometrical applications, clustering algorithms should be able to determine clusters of arbitrary geometrical shape, clusters of non-uniform density, and clusters belonging to a low-dimensional hyperspace. The algorithms should be robust even in the case when significant amount of noise is present. In some applications it is necessary to cluster according to the local conditions only. For geometrical applications and classification problems it is often important not only to determine clusters (as subsets of data points) but also to draw an accurate boundary between close neighboring clusters. For real-time applications, it is also very important that clustering algorithms work sufficiently fast, preferably with linear complexity with respect to the number of data points. Therefore, there is a strong demand for development of new efficient and robust clustering algorithms.

Existing clustering approaches proposed in last 15 years include partitioning methods (K-MEANS and K-MEDOIDS) [4,5], hierarchical methods [6,7], density-based methods [8–10], model-based methods [11,12], graph-based methods [13,14], grid-based methods [15,16], etc., and their various combinations and improvements [17–20], etc. Some of these methods became classical. They proved to be successful in detecting certain cluster structures.

Among well-known clustering algorithms there are PAM (partitioning around medoids) [5], CLARANS (partitioning, medoid-based) [21], BIRCH (hierarchical) [7], OPTICS (hierarchical) [22], CHAMELEON (hierarchical) [23], DBSCAN (density-based) [8], DENCLUE (density-based) [9], CLIQUE (density-based and grid-based) [15], CURE (hierarchical) [6], AMOEBA (graph-based) [13], AUTOCLUST

* Corresponding author. Tel.: +65 67905442.
E-mail address: eosourina@ntu.edu.sg (O. Sourina).

0031-3203/$30.00
doi:10.1016/j.patcog.2008.01.021

(graph-based) [14], EM algorithm [24], MCLUST (model-based) [25], etc.

Most of the clustering methods require setting of the user specified parameters or prior knowledge to produce their best results. For example, partitioning methods such as K-MEANS and K-MEDOIDS face the problem of prescribing the number of clusters in advance. Classic hierarchical methods, such as single-linkage and complete-linkage methods, require setting of the merge/split conditions to end the clustering process. The representative density-based methods DBSCAN and DENCLUE need to set density threshold parameter. Some other methods require specific data points distribution [26].

Except for specific applications when we have complete knowledge about the data set to ensure the validity of chosen parameters, these non-automatic methods are quite unstable because of the probability of introducing human-generated bias and are not efficient because of time-consuming procedure of parameters tuning. So, it is important to develop a clustering method which can perform data clustering automatically.

Another new requirement to clustering algorithms is discovering of arbitrary shape clusters. Pure partitioning methods absolutely lack the ability to deal with clusters of arbitrary shape. Hierarchical methods can detect clusters of relatively complicated, but still not arbitrary, only convex shape. Even improved hierarchical algorithms such as CURE cannot detect clusters of very complicated shape. Although hierarchical clustering can be effective in knowledge discovery, the cost of creating dendrogram could be prohibitively expensive since the corresponding algorithms are at least quadratic with respect to the number of data points [9]. Arbitrary-shaped clusters can be detected by density-based algorithms such as DBSCAN and DENCLUE. The basic assumption in DBSCAN is that each data point inside cluster must have at least the certain number of data points in its neighborhood of the given radius. The primary idea here is that density of the clusters should exceed some global threshold [8]. But this global threshold assumption prevents DBSCAN to determine clusters of different density. DENCLUE, contrary to DBSCAN exploits the idea of field function which efficiently measures local density and results in a fast performance. Still, it has the similar difficulty in recognition the clusters of different density. The reason is the same—dependence upon a predefined global threshold.

The distinguishing of different in density clusters is performed (to different extent) in modern graph-based algorithms such as CHAMELEON, AMOEBA and AUTOCLUST. Among them AUTOCLUST seems to be the most effective one. It is able to distinguish not only clusters of different density, but also sparse clusters which are adjacent to high-density clusters. But all algorithms cannot recognize clusters with significantly non-uniform internal density which appear in computer graphics applications (for example, in point-based graphics), pattern recognition, and geo-image processing. In Fig. 1, two examples of non-uniform density clusters are given. Just by visual inspection, we can easily recognize those clusters whose density is gradually changing inside, since the variation of discrimination according to the variation of density is natural and important. Moreover, the complexity of all these algorithms exceeds $O(N)$ because they require the computation of Delaunay Diagrams of similar graphs. The total complexity is estimated as $O(N \log N)$, where N is the number of data points. Graph-based algorithms are efficient in spatial databases and geographical information systems (GIS), but their application to computer geometry, for example, is limited due to non-linear performance and necessity of total re-computation when the area of clustering is dynamically changing.

Another problem arising from computer geometry clustering is detection of clusters belonging to low-dimensional hyperspaces. This problem becomes even more important due to recent development of point-based graphics. It is associated with the problem of surfels [27] determination for complex-shaped surfaces. When a surface is defined by the cloud of points unrelated to each other (for example, obtained from three-dimensional scanning process), for a given point it is necessary to determine its neighborhood consisting of such data points, which could be connected to it by a sufficiently short path along the surface. Such determination requires accurate clustering in Euclidean vicinity of a given data point taking in account that such vicinity can contain several different surfels and/or noise.

In common clustering algorithms, boundary detection is usually not incorporated. But, as it was mentioned above, the problem of cluster boundary definition arises, for example, in computer geometry applications where we need to compute (and draw) an accurate boundary of a geometric object determined by the set of points. In work [28], a definition of cluster as a solid described by a set of points endowed with influence functions was introduced. Such definition is capable not only to describe granular property of a cluster but also its boundary.

Classification problems give us another example when boundary detection of clusters is really important. Knowledge of cluster boundaries makes it possible to classify new data without repeating the clustering process [29]. Thus, detection of cluster boundaries is necessary for reasoning about clusters [30].

Furthermore, for clustering problems, detection of noise or outliers is also an important task. In some applications, such as point-based graphics, the definitions of noise and outliers are different. Noise is caused by environmental factors and is not part of data. Outliers are the original data points which do not belong to any cluster. In this paper, we consider outliers the same as noise. Many traditional methods can detect noise when the density of clusters does not vary greatly. But it becomes quite a challenge to detect noise when the density of clusters changes not only between clusters but also inside clusters.

In this paper, we propose a new algorithm ADACLUS (ADAptive CLUStering) based on local-adaptive influence function. It allows to automatically discover clusters of arbitrary shape and different density (density can be different inside cluster as well as between clusters), to detect cluster boundaries, and it is robust to noise. Contrary to many existing algorithms, ADACLUS does not require any parameter pre-setting if it works in pure automatic mode. But it can also work in manual mode when it gives the user a possibility to tune clustering process by setting three parameters all of