Pattern Recognition 45 (2012) 3017–3033

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

A robust adaptive clustering analysis method for automatic identification of clusters

P.Y. Mok*, H.Q. Huang, Y.L. Kwok, J.S. Au

Institute of Textiles and Clothing, The Hong Kong Polytechnic University, Hunghom, Hong Kong

ARTICLE INFO

Article history:
Received 4 January 2010
Received in revised form
11 December 2011
Accepted 2 February 2012
Available online 16 February 2012

Keywords:
Cluster analysis
Cluster validity
Fuzzy clustering
Fuzzy C-Means
Cluster ensembles

ABSTRACT

Identifying the optimal cluster number and generating reliable clustering results are necessary but challenging tasks in cluster analysis. The effectiveness of clustering analysis relies not only on the assumption of cluster number but also on the clustering algorithm employed. This paper proposes a new clustering analysis method that identifies the desired cluster number and produces, at the same time, reliable clustering solutions. It first obtains many clustering results from a specific algorithm, such as Fuzzy C-Means (FCM), and then integrates these different results as a judgement matrix. An iterative graph-partitioning process is implemented to identify the desired cluster number and the final result. The proposed method is a robust approach as it is demonstrated its effectiveness in clustering 2D data sets and multi-dimensional real-world data sets of different shapes. The method is compared with cluster validity analysis and other methods such as spectral clustering and cluster ensemble methods. The method is also shown efficient in mesh segmentation applications. The proposed method is also adaptive because it not only works with the FCM algorithm but also other clustering methods like the *k*-means algorithm.

1. Introduction

Cluster analysis aims to partition a large number of data into different subsets or groups so that the requirements of homogeneity and heterogeneity are fulfilled. Homogeneity requires that data in the same cluster should be as similar as possible and heterogeneity means that data in different clusters should be as different as possible [1]. Typical clustering activity involves three sequential steps [2]: data/object representation, definition and computation of data proximity, and clustering/grouping, as shown in Fig. 1. Data/object representation refers to problem definition including the number of clusters, the number of available data, and the number, type, and scale of the data variables available to the clustering algorithm. Data proximity, also known as inter-object similarity, is usually measured by a distance function defined on pairs of data. A variety of distance measures are in use for different purposes. The grouping step can be performed in a number of ways, for instance hierarchical approach, partitional approach and other algorithms can be employed. Cluster analysis is widely used in areas such as market research, pattern recognition [3], image segmentation [4], and mesh segmentation [5].

Different clustering algorithms have been developed in the past, and some examples are shifting grid [6], SOFM neural networks [7] and Evidential C-Means [8]. The availability of such a vast collection of clustering algorithms in the literature can easily confuse users attempting to select algorithm for a specific problem. When presented with data, all clustering algorithms will produce clusters regardless of whether the data contain clusters or not. There is no clustering technique that is universally applicable in uncovering the variety of structures present in multidimensional data sets [2]. It is because clustering algorithms often contain implicit assumptions about cluster shape and grouping criteria used. Clustering algorithms must be carefully selected by evaluating (1) the manner in which clusters are formed, (2) the structure of the data, and (3) sensitivity of the clustering technique [2].

Although clustering is a useful and challenging problem with great potential in applications, its application must be cautiously handled. Otherwise, the technique can easily be abused or misapplied. Cluster number and similarity measure are the two most important assumptions of clustering analysis, which affect the overall quality of the results. In most of the automatic clustering algorithms, the cluster number must be first defined. This is true for most popular algorithms like the Fuzzy C-Means (FCM) clustering algorithm. Some researchers [9–13] have proposed cluster validity indices to validate the cluster results so as to obtain the optimal cluster number. Apart from identifying the

* Corresponding author. Tel.: +852 2766 4442; fax: +852 2773 1432.
E-mail address: tracy.mok@inet.polyu.edu.hk (P.Y. Mok).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2012.02.003

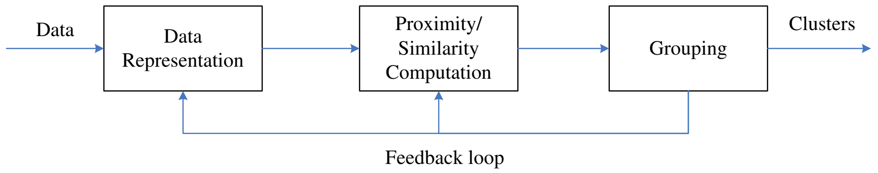

Fig. 1. Steps in clustering.

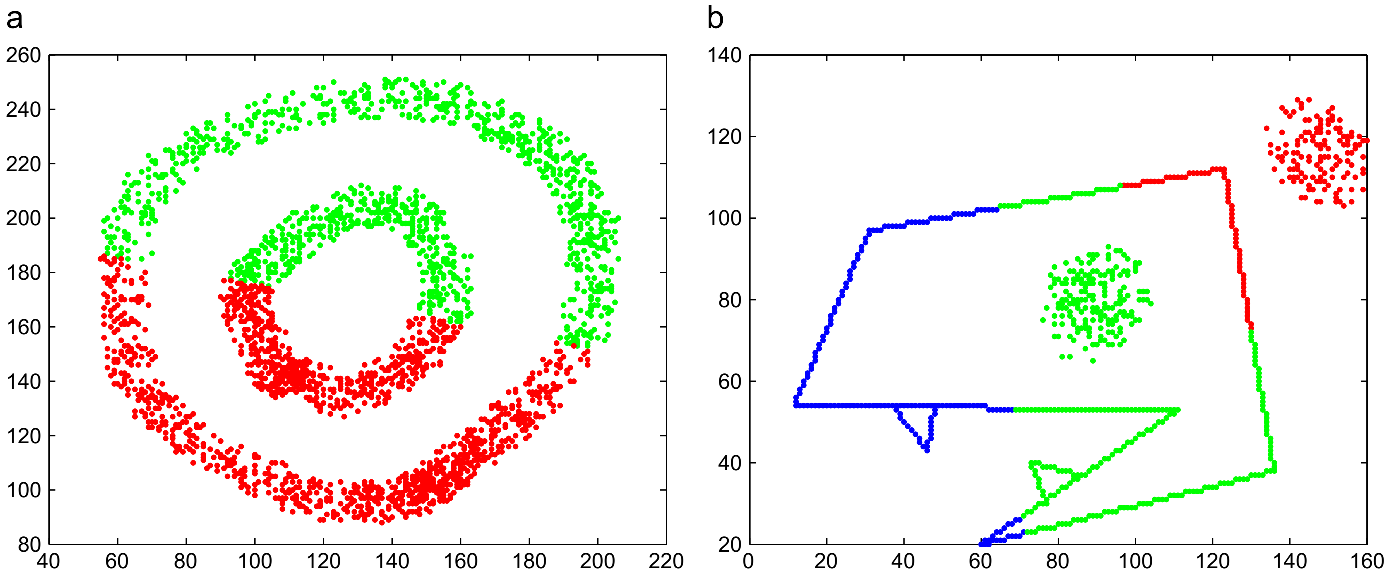

Fig. 2. Clustering data sets by Fuzzy C-Means clustering algorithm. (a) cluster number=2 and (b) cluster number=3.

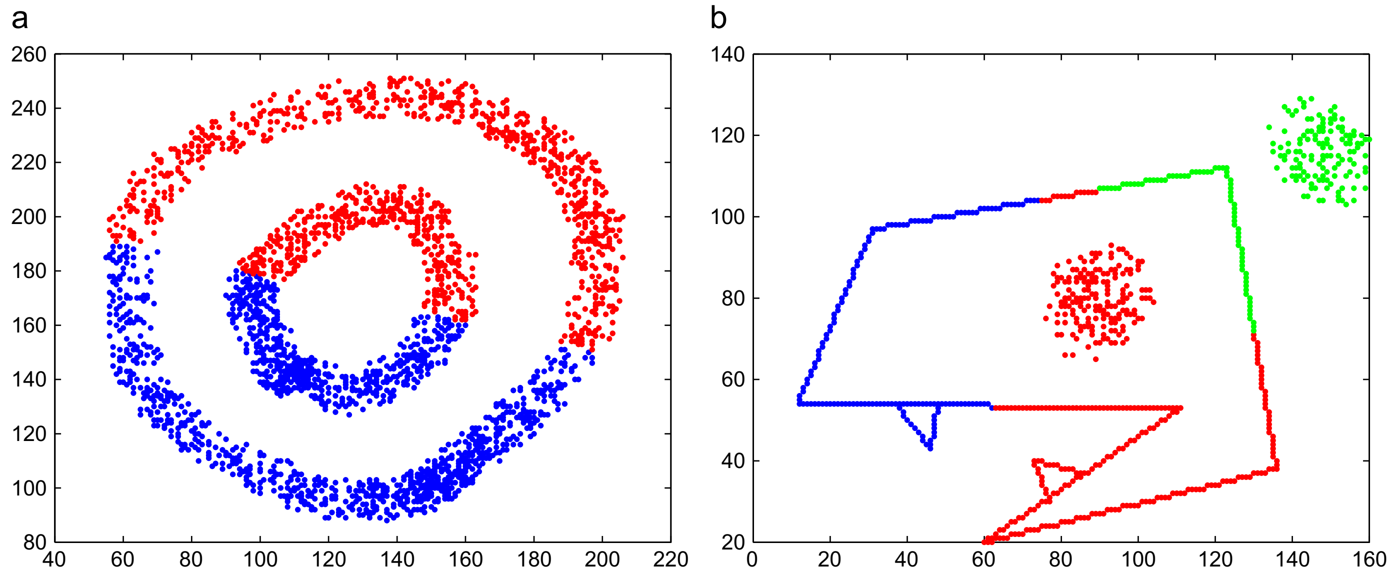

Fig. 3. Clustering data sets by *k*-means clustering algorithm. (a) cluster number=2 and (b) cluster number=3.

optimal cluster number, effective clustering requires the algorithm to be robust for data sets of different shapes. Sometimes, correct cluster numbers do not guarantee that a data set can be properly partitioned in the desired way. Most widely used clustering algorithms assumed distance based similarity measures [2], upon which the grouping process is carried out. There are varied types of distance based similarity measures, such as Euclidean distance, Manhanttan distance, and Mahalanobis distance. The similarity measure must be chosen carefully. For instance, as shown in Figs. 2 and 3, data sets are not well partitioned by either FCM or *k*-means algorithms, even though the correct cluster numbers are given. FCM and *k*-means algorithms use centroid-based distance as similarity measurement. In Figs. 2 and 3, different clusters are depicted in different colours, and all figures hereafter follow the same colour scheme to illustrate cluster results.

The objective of this paper is to propose a robust and adaptive clustering analysis method that produces reliable clustering