Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition 38 (2005) 1887–1901
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

A new approach to clustering data with arbitrary shapes

Mu-Chun Su*, Yi-Chun Liu

Department of Computer Science & Information Engineering, National Central University, Chung-Li, Taiwan, R.O.C.

Received 15 July 2004; received in revised form 8 March 2005; accepted 22 April 2005

Abstract

In this paper we propose a clustering algorithm to cluster data with arbitrary shapes without knowing the number of clusters in advance. The proposed algorithm is a two-stage algorithm. In the first stage, a neural network incorporated with an ART-like training algorithm is used to cluster data into a set of multi-dimensional hyperellipsoids. At the second stage, a dendrogram is built to complement the neural network. We then use dendrograms and so-called tables of relative frequency counts to help analysts to pick some trustable clustering results from a lot of different clustering results. Several data sets were tested to demonstrate the performance of the proposed algorithm.

Keywords: Cluster analysis; ART; Clustering; Unsupervised learning; Hierarchical partitioning

1. Introduction

Clustering algorithms are effective tools for exploring the structures of complex data sets, and therefore, are of great value in a number of applications [1–6]. For most of clustering algorithms, two crucial problems required to be solved are (1) the determination of the optimal number of clusters and (2) the determining of the similarity measure based on which patterns are assigned to corresponding clusters. The estimation of the number of clusters in the data set is the so-called cluster validity problem. Conventional approaches to solving the cluster validity problem usually involve increasing the number of clusters, and/or merging the existing clusters, computing some certain cluster validity measures in each run, until partition into optimal number of clusters is obtained [7–13]. Since most validity measures usually impose a certain structure on the data, these approaches fail to estimate the correct number of clusters in real data with a large variety of distributions within and between clusters. The second crucial problem encounters a similar situation as the first problem does. While it is easy to consider the idea of a data cluster on a rather informal basis, it is very difficult to give a formal and universal definition of a cluster. Most of the conventional clustering methods assume that patterns having similar locations or constant density create a single cluster. In order to mathematically identify clusters in a data set, it is usually necessary to first define a measure of similarity or proximity which will establish a rule for assigning patterns to the domain of a particular cluster center. As it is to be expected, the measure of similarity is problem dependent. That is, different similarity measures will result in different clustering results.

Lately, neural networks have been used for data clustering. Similar to the *k*-means algorithm, the winner-take-all network suffers the disadvantage of the requirement of the number of clusters to be created. On the other hand, the adaptive resonance theory (ART) networks are able to cluster data without pre-specifying the number of clusters [14–16]. To cope with the problem of estimating the number of clusters, the ART networks adopt a so-called vigilance parameter with which the networks decide when to create

* Corresponding author. Tel.: +886 3 422 7151; fax: +886 3 422 2681.
E-mail address: muchun@csie.ncu.edu.tw (M.-C. Su).

0031-3203/$30.00
doi:10.1016/j.patcog.2005.04.010

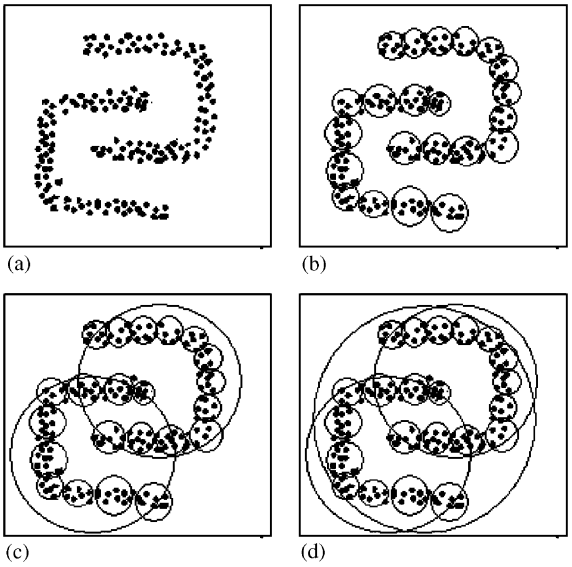

Fig. 1. A data set consisted of two horseshoe-shaped clusters: (a) the original input data; (b) 22 clusters are formed in the first ART network using a stricter distance measure; (c) 22 clusters are merged into two larger clusters in the second ART network using a medium distance measure; (d) a large cluster is finally formed in the third ART network using a looser distance measure.

new clusters. Although the ART networks do not explicitly specify the number of clusters, the vigilance parameter to some extent implicitly pre-specifies it. That is, the larger the value of the vigilance parameter is the larger the number of clusters we have.

Several approaches have been proposed to tackle one or both of the aforementioned problems [17–26] (to name just a few here). For example, in Ref. [17], Eltoft et al. proposed a new unsupervised neural network which is capable of clustering a set of data according to a given genetic inter-point similarity measure without knowing in advance the number of clusters to be created. The proposed neural network consists of a two-layer feedforward neural network and a threshold calculating unit. In addition, they used the inverse Euclidean distance as a measure of similarity. A new recursive algorithm for hierarchical fuzzy partitioning is proposed in Ref. [18]. The algorithm has the advantages of hierarchical clustering, while maintaining fuzzy clustering rules. Several different approaches to estimating the number of clusters in a data set is to interpret self-organizing feature maps trained by the data sets [19–23]. In Refs. [24,25], a kind of "point symmetry distance" was proposed to group a given data set into a set of clusters of different geometrical structures. A new method, based on the principle of data induced metric (DIM), was developed for partitional clustering of nonconvex clusters of arbitrary shape [26].

In this paper, we propose a new clustering algorithm which can be used for partitioning well-separated nonconvex clusters, the geometry of which cannot be analytically described, without knowing the number of clusters in advance. The remaining of this paper is organized as follows. In Section 2 the proposed algorithm is discussed. Section 3 presents the experimental results. In Section 4, an auxiliary method for the proposed algorithm is introduced. Finally, Section 5 concludes the paper.

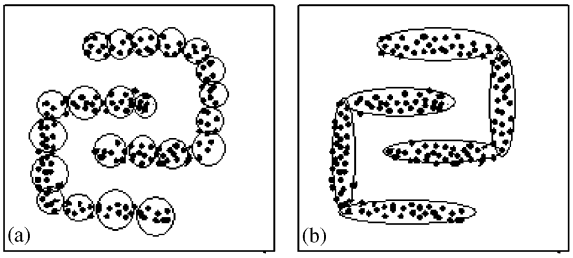

Fig. 2. The idea of using sets of substructures to approximate clusters of arbitrary shapes: (a) 22 circles are used; (b) 6 ellipses are used.

2. The proposed clustering algorithm

Owing to many appealing properties, ART networks provide a natural basis for many researchers. However, a crucial question may be immediately raised about the ART networks. How much confidence do we have on the clustering results created by ART networks since the results are dependent on the values of the vigilance parameter? It seems that we need a complementary method to help us to choose a more trustable clustering result among many suspicious ones created by ART networks using different values of the vigilance parameter.

Hierarchical clustering seems to provide an appealing solution to this question. Hierarchical clustering method creates a hierarchical decomposition of a data set. The algorithm can be agglomerative or divisive. An agglomerative hierarchical algorithm starts with the disjoint clustering, which places each of the objects in an individual cluster. Then it gradually merges these atomic clusters into larger and larger clusters until all data are in a single cluster. A divisive hierarchical algorithm performs the task in the reverse order. The hierarchical clustering is usually represented by a dendrogram which consists of layers of nodes, each representing a cluster. Lines connect nodes representing clusters that are nested into another. Each level of the dendrogram represents a clustering of the data set.

In fact, hierarchical ART architectures have already been proposed to extend the knowledge representation capabilities of single ART networks [27–30]. Basically, they all employ a cascade of ART networks. Each ART network receives the prototypes (or the difference between the input and the prototypes) formed in the previous ART network. The value of the vigilance parameter for each ART network is then gradually decreased. Therefore, many small clusters are