Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 40 (2007) 1425–1431
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

A clustering algorithm based on maximal θ-distant subtrees

Li Yujian*

Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology, College of Computer Science and Technology, Beijing University of Technology, Beijing 100022, China

Received 13 December 2005; received in revised form 15 January 2006; accepted 5 October 2006

Abstract

This paper presents a clustering algorithm based on maximal θ-distant subtrees, the basic idea of which is to find a set of maximal θ-distant subtrees by threshold cutting from a minimal spanning tree and merge each of their vertex sets into a cluster, coupled with a post-processing step for merging small clusters. The proposed algorithm can detect any number of well-separated clusters with any shapes and indicate the inherent hierarchical nature of the clusters present in a data set. Moreover, it is able to detect elements of small clusters as outliers in a data set and group them into a new cluster if the number of outliers is relatively large. Some computer simulations demonstrate the effectiveness of the clustering scheme.

Keywords: Maximal θ-distant subtree; Minimal spanning tree; Clustering algorithm; Threshold cutting; Number of clusters

1. Introduction

Clustering is used to group similar objects into clusters, it has wide applications in speech and image processing, biological information computing and data mining, etc. [1–5]. Considering a data set $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, the goal of clustering process is to find a partition composed of such subsets $\{V_k\}_{k=1}^{p}$ so that $X = \bigcup_{k=1}^{p} V_k, \forall 1 \leqslant i, j, k \leqslant p, i \neq j, V_i \bigcap V_j = \phi, V_k \neq \phi$, where all V_k are called clusters and any two elements in a same cluster should be more similar than those in different clusters. Classical clustering techniques include K-means algorithm, hierarchical agglomerative clustering algorithms and the minimal spanning tree-based (MST-based) algorithms [6,7]. There have also been many recent clustering methods such as CURE [8], Chameleon [9], hierarchical growing cell structures [10], highly connected subgraphs [11], generality-based conceptual clustering [12], relative neighborhood graphs [13] and so on. Most of the clustering algorithms have to specify some parameters in advance. For example, the K-means algorithm requires the a priori specification of the number of clusters, which may not be feasible in many situations. In fact, if no pertinent knowledge or experience is given, it is often difficult to specify a parameter for any clustering technique. How to avoid or reduce the difficulty remains to be an important problem in the field of pattern recognition and artificial intelligence [14].

The motivation of this work is to develop a good clustering algorithm requiring only simple specification of parameters by improving the minimal spanning tree (MST)-based algorithms. The basic idea of the classical MST-based method is to partition a data set into clusters by cutting inconsistent edges from a MST. However, the definition of inconsistency is problem specific, sometimes it requires the knowledge about the shape of the clusters [15]. The limitations of inconsistent edges can be partly overcome by a new MST-based method [16], which is also called the "long-edge cutting" method in this paper because its basic idea is to partition a data set into clusters by cutting long edges from an MST. One obvious drawback of the long-edge cutting method is that it cannot directly determine how many clusters there should be in a data set, although it is able to automatically select a good number of clusters by examining the optimal K-clustering for all $K = 1, 2, \ldots$, up to some large number.

* Tel.: +86 10 67392879 2514.
E-mail address: liyujian@bjut.edu.cn.

0031-3203/$30.00
doi:10.1016/j.patcog.2006.10.003

We intend to investigate the problem of how to directly determine the reasonable number of clusters in a data set. The new idea we use here is "threshold cutting" which means "cutting all edges whose lengths are greater than a certain threshold $\theta \geqslant 0$". We have shown that all subtrees generated by threshold cutting from a MST are maximal θ-distant subtrees, the vertex sets of which exactly form a unique partition of the data set without considering their orders. Based on this important property, we have developed a new clustering algorithm—MDS_CLUSTER—that is able to detect non-overlapping clusters of any shape requiring only simple specification of one parameter, namely, the least number of elements in each cluster. MDS_CLUSTER is also able to provide several main levels of clusters in a hierarchy which is different from the all-level hierarchy generated by the traditional hierarchical agglomerative clustering scheme. In addition, MDS_CLUSTER is able to detect outliers in a data set and group them into a background cluster if the number of outliers is relatively large.

In this paper, we discuss some properties of MSTs and maximal θ-distant subtrees in Section 2, present the clustering algorithm MDS_CLUSTER in Section 3, describe computer simulations in Section 4 and make conclusions in Section 5.

2. Some properties of MSTs and maximal θ-distant subtrees

Considering a data set $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ and a distance function ρ defined on $X \times X$, we define the induced graph of X with respect to ρ as a weighted (undirected) graph $G_\rho(X) = (V, E)$, where the vertex set $V = X$, and the edge set $E = \{(\mathbf{x}_i, \mathbf{x}_j) |$ for $\mathbf{x}_i, \mathbf{x}_j \in X$, and $i \neq j\}$. Hence, $G_\rho(X)$ is a complete graph. Each edge $(u, v) \in E$ has a length or weight of the distance $\rho(u, v)$ between u and v, which could be defined as the Euclidean distance, Manhattan distance or some other distance, had better but may not be a metric.

A spanning tree T of $G_\rho(X)$ is a connected subgraph of $G_\rho(X)$ such that T contains every vertex of $G_\rho(X)$ and T does not contain any circle. An MST is a spanning tree with the minimum total distance. If $T = (V_T, E_T)$, where V_T and E_T are, respectively, the vertex set and the edge set of T, then the total distance of T can be defined as the total sum of all its edge lengths, namely, $\rho(T) = \sum_{e \in E_T} \rho(e)$.

Let ϕ represent the empty set and define $\max \phi = 0$, $\min \phi = +\infty$. A maximal θ-distant subtree of $G_\rho(X)$ is defined as a subtree $T = (V_T, E_T)$ of $G_\rho(X)$ which satisfies the following three conditions:

(1) T is an MST of the subgraph $G_\rho(V_T)$;
(2) $\max\{\rho(e) | e \in E_T\} \leqslant \theta$;
(3) $\min\{\rho(e) | e \in E[V_T, \overline{V}_T]\} > \theta$ where $\overline{V}_T = X - V_T$ and $E[V_T, \overline{V}_T]$ represents the set of all edges that have one vertex in V_T and the other vertex in $\overline{V}_T$.

If $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is a data set with a distance ρ, an MST or a maximal θ-distant subtree of X is defined as that of $G_\rho(X)$. For example, the tree in Fig. 1b is an MST of the data set in Fig. 1a, and the trees in Fig. 2a and b are two different levels of maximal θ-distant subtrees in the data set, where ρ is the Euclidean distance. It is not difficult to find that there may be more than one MST and several different levels of maximal θ-distant subtrees for a data set. Theorems 1–3 describe some important properties of an MST or a maximal θ-distant subtree.

Theorem 1. *Suppose that T_1 and T_2 are two MSTs of $G_\rho(X)$, if all edges of T_1 and that of T_2 are enumerated, respectively, as $e_1^1, e_2^1, \ldots, e_m^1$ and $e_1^2, e_2^2, \ldots, e_m^2$, there must exist a permutation $j_1, j_2, \ldots, j_m$ of $1, 2, \ldots, m$ such that $\rho(e_k^1) = \rho(e_{j_k}^2)$, $k = 1, 2, \ldots, m$.*

Proof. If $T_1 = T_2$, Theorem 1 is obviously true; otherwise, there must be an edge $e \in T_1 \backslash T_2$ such that $T_2 + e$ contains only one cycle $C(e)$ which contains at least one edge $e' \in T_2 \backslash T_1$. It can be shown that $\rho(e) = \rho(e')$.

If $\rho(e) < \rho(e')$, the spanning tree $T' = T_2 + e - e'$ has a total distance such that $\rho(T') < \rho(T_2)$, this contradicts that T_2 is an MST. Hence $\rho(e) \geqslant \rho(e')$ must be satisfied.

If $\rho(e') < \rho(e)$, the spanning tree $T'' = T_1 + e' - e$ has a total distance such that $\rho(T'') < \rho(T_1)$, this contradicts that T_1 is an MST. Hence $\rho(e') \geqslant \rho(e)$ must be satisfied.

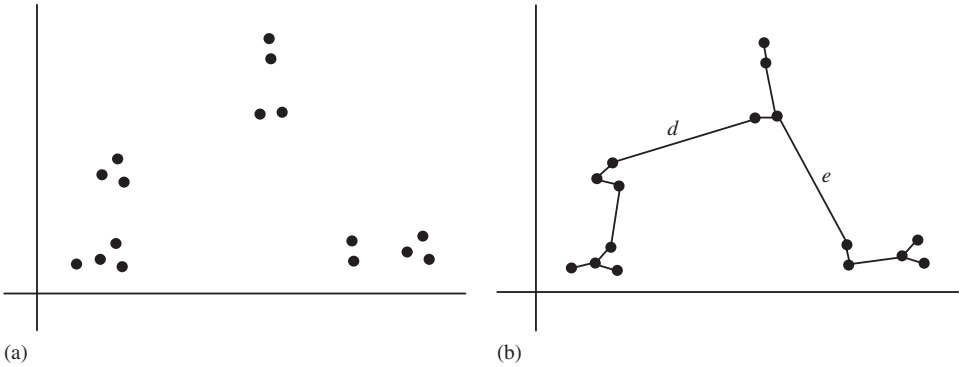

Fig. 1. (a) A data set X composed of 2D points. (b) An MST of X.