Information Sciences 272 (2014) 29–48

Contents lists available at ScienceDirect

Information Sciences

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

A general framework of hierarchical clustering and its applications

Ruichu Cai [a,*], Zhenjie Zhang [b], Anthony K.H. Tung [c], Chenyun Dai [d], Zhifeng Hao [a]

[a] *Faculty of Computer Science, Guangdong University of Technology, Guangzhou, PR China*
[b] *Advanced Digital Sciences Center, Illinois at Singapore pte, Singapore, Singapore*
[c] *School of Computing, National University of Singapore, Singapore*
[d] *Department of Computer Science, Purdue University, USA*

ARTICLE INFO

Article history:
Received 16 February 2012
Received in revised form 26 December 2013
Accepted 9 February 2014
Available online 20 February 2014

Keywords:
Clustering
Hierarchical
k-Means
k-Median
Streaming algorithm

ABSTRACT

Hierarchical clustering problem is a traditional topic in computer science, which aims to discover a consistent hierarchy of clusters with different granularities. One of the most important open questions on hierarchical clustering is the identification of the meaningful clustering levels in the hierarchical structure. In this paper, we answer this question from algorithmic point of view. In particular, we derive a quantitative analysis on the impact of the low-level clustering costs on high level clusters, when agglomerative algorithms are run to construct the hierarchy. This analysis enables us to find meaningful clustering levels, which are independent of the clusters hierarchically beneath it. We thus propose a general agglomerative hierarchical clustering framework, which automatically constructs meaningful clustering levels. This framework is proven to be generally applicable to any k-clustering problem in any α-relaxed metric space, in which strict triangle inequality is relaxed within some constant factor α. To fully utilize the hierarchical clustering framework, we conduct some case studies on k-median and k-means clustering problems, in both of which our framework achieves better approximation factor than the state-of-the-art methods. We also extend our framework to handle the data stream clustering problem, which allows only one scan on the whole data set. By incorporating our framework into Guha's data stream clustering algorithm, the clustering quality is greatly enhanced with only small extra computation cost incurred. The extensive experiments show that our proposal is superior to the distance based agglomerative hierarchical clustering and data stream clustering algorithms on a variety of data sets.

1. Introduction

Clustering analysis is a well studied topic in computer science [14,16,3,31,2,11,10,5,41]. Generally speaking, clustering analysis tries to divide the unlabelled objects into several groups, maximizing the similarities among objects in the same group while minimizing the similarities among objects from different groups. It is widely used in many real applications, such as market analysis, image segmentation and information retrieval. While traditional clustering techniques usually just

[*] Corresponding author. Tel.: +86 015800030523; fax: +86 20 39323163.
E-mail addresses: cairuichu@gmail.com (R. Cai), zhenjie@adsc.com.sg (Z. Zhang), atung@comp.nus.edu.sg (A.K.H. Tung), daic@cs.purdue.edu (C. Dai), mazfhao@scut.edu.cn (Z. Hao).

http://dx.doi.org/10.1016/j.ins.2014.02.062
0020-0255/© 2014 Elsevier Inc. All rights reserved.

split the objects based on a specified or estimated cluster number, hierarchial clustering [13,34] aims to construct a hierarchical structure consisting of clusters with different granularities.

The interests on hierarchical clustering stem from different applications. First, it is well observed that people understand the universe in a hierarchical manner. In zoology, for example, gorilla and chimpanzee are all animals similar to human given a high level of species categorization, while both of them are quite different from human beings when zooming into the specific category of "Euarchonta". To better understand the relationships among unknown objects, it is necessary and fundamental to construct a hierarchical clustering rather than clustering with a single granularity, e.g. recovering the hierarchy of natural topics in text mining [43,42,25]. Second, hierarchical clustering is useful in many operational tasks. In sensor network, a well designed hierarchical clustering on the sensor nodes is able to improve the structure of the network system [8], leading to less communication cost and more energy savings on the nodes. Third, a good hierarchical clustering provides concise summarizations of the data on different granularities. These summarizations facilitate applications in scenarios with strict memory constraints, such as data streams. Existing clustering algorithms on data stream usually exploit the hierarchical structure for fast and accurate clustering on large data set [17,12,37].

In this paper, we focus on the general k-clustering problem, which discovers k *centers* in the space, minimizing the *clustering cost*, i.e. the sum of the distances from the data points to the nearest center. Given a point group of size n, the standard hierarchical clustering is a natural extension of k-clustering problem, constructing level-wise consistent k-clusters with k from 1 to n on different levels. Specifically, each clustering level L_i is the refinement on the level L_{i-1}, with L_1 is exactly the original data set. In Fig. 1, we present an example of hierarchical clustering on 1-dimensional data. It is straightforward to verify that clustering on level L_i simply merges two centers in the clustering on level L_{i-1}. In the last decade, extensive efforts were devoted to k-clustering problems with respect to a wide spectrum of distance functions, such as squared Euclidean distance (k-means) [3,31] and general metric distance (k-median) [2,11,10,5], leading to algorithms achieving constant approximations on the clustering cost. However, solutions to hierarchical k-clustering problem with performance guarantee were not available until recently [13,38,34], due to the hardness on the approximation requirements on all levels. While existing hierarchical clustering algorithms only return results with large approximation factors, an important question arises on the meaningfulness of the hierarchy with all clustering levels of all possible cluster numbers, since most of the levels do not provide additional categorization information to other levels. In this paper, we address this problem by carefully identifying more important clustering levels from the hierarchy. Intuitively, a clustering level L_i may be more informative, if the levels beneath it in the hierarchy do not affect the clustering cost of L_i achieved by the clustering algorithm. Later, we will give detailed analysis of this intuition using a Chinese Restaurant Process model [9]. This criterion distinguishes core clustering levels from trivial ones. Take Fig. 1 an example, L_2 is simply trivial compared with L_1, while L_4 may give a much better abstraction on the four clusters in the original data set.

To effectively and efficiently discover a clustering hierarchy containing only important and meaningful clustering levels, we propose a general agglomerative hierarchical clustering framework. This framework is general enough to handle any k-clustering problem in any α-relaxed metric space, in which the strict triangle inequality is relaxed with some constant factor α. Given a k-clustering algorithm and the relaxed metric space, the framework constructs the hierarchical clustering in a bottom-up manner. In each construction iteration, the framework first selects the appropriate size s_i for the next clustering level to build. The clustering algorithm is then invoked to find s_i centers in the space as elements in the new level. The construction process terminates when reaching the top level with exactly one center. If running the framework on the 1-dimensional data in Fig. 1, it skips the first two levels L_2 and L_3, and directly jumps to level L_4. Another level L_5 is selected and constructed in next round, before L_7 capping the clustering hierarchy.

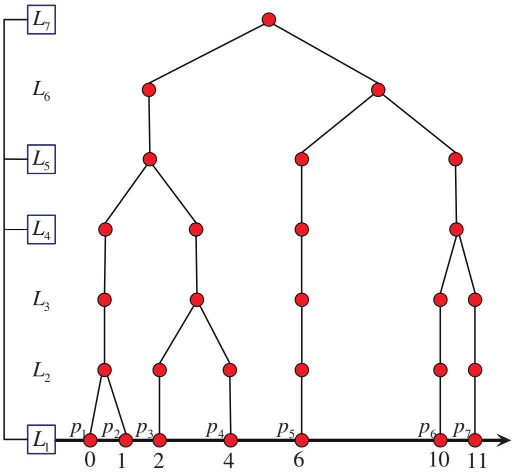

Fig. 1. Example of hierarchical clustering on 1-dimensional data.