Chapter 14
Data Field for Hierarchical Clustering

Shuliang Wang
The University of Pittsburgh, USA & Wuhan University, China

Wenyan Gan
Nanjing University of Science and Technology, China

Deyi Li
Tsinghua University, China

Deren Li
Wuhan University, China

ABSTRACT

In this paper, data field is proposed to group data objects via simulating their mutual interactions and opposite movements for hierarchical clustering. Enlightened by the field in physical space, data field to simulate nuclear field is presented to illuminate the interaction between objects in data space. In the data field, the self-organized process of equipotential lines on many data objects discovers their hierarchical clustering-characteristics. During the clustering process, a random sample is first generated to optimize the impact factor. The masses of data objects are then estimated to select core data object with nonzero masses. Taking the core data objects as the initial clusters, the clusters are iteratively merged hierarchy by hierarchy with good performance. The results of a case study show that the data field is capable of hierarchical clustering on objects varying size, shape or granularity without user-specified parameters, as well as considering the object features inside the clusters and removing the outliers from noisy data. The comparisons illustrate that the data field clustering performs better than K-means, BIRCH, CURE, and CHAMELEON.

INTRODUCTION

The rapid advance in massive data acquisition, transmission and storage results in the growth of vast computerized datasets at unprecedented rates. These datasets come from various sectors, e.g., business, education, government, scientific community, Internet, or one of many readily available off-line and online data sources in the form of text, graphic, image, video, audio, animation, hyperlinks, markups, and so on (Li, Zhang, & Wang, 2006; Bhatnagar, Kaur, & Mignet, 2009). Moreover, they are continuously increasing and amassed in both attribute depth and scope of in-
stances every time. Although many decisions are made on large datasets, the huge amounts of the computerized datasets have far exceeded human ability to completely interpret (Li et al., 2006). In order to understand and make full use of these data repositories when making decisions, it is necessary to develop some technique for uncovering the physical nature inside such huge datasets.

Clustering is one of the techniques to discover a segmentation rule from these data repositories. It assigns a set of objects into clusters (subsets) by virtue of their observations so that objects are similar to one another within the same cluster and are dissimilar to the objects in other clusters (Murtagh, 1983; Grabmeier & Rudolph, 2002; Xu & Wunsch, 2005; Li, Wang, & Li, 2006; Malik et al., 2010). It is an unsupervised technique without the knowledge what causes the grouping and how many groups exist (Song, Hu, & Yoo, 2009; Engle & Gangopadhyay, 2010; Silla & Freitas, 2011). The arbitrary shaped clustering was further treated (Wan, Wang, & Su, 2010). Clustering may be implemented on hierarchy, partition, density, grid, constraint, subspace and so on (Sander et al., 1998; Kwok et al., 2002; Grabmeier & Rudolph, 2002; Parsons, Haque, & Liu, 2004; Zhang et al., 2008; Horng et al., 2011).

### Hierarchy-Based Clustering

Hierarchy-based clustering finds successive clusters using previously established clusters (Murtagh, 1983). It uncovers a nested sequence of clusters with a single, all-inclusive cluster at the top and single-point clusters at the bottom. In the sequence, each cluster is nested into the next cluster (Guha, Rastogi, & Shim, 1998). Hierarchical clustering algorithms are either agglomerative or divisive. Agglomerative algorithms start with each element as a disjoint set of clusters and merge them into successively larger clusters (Sembiring, Zain, & Embong, 2010). Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters (Malik et al., 2010). One of the simplest methods is K-means that take each point belonging to a given data set and associate it to the nearest centroid after each cluster is defined with a centroid on the basis of a sample (MacQueen, 1967). But its cluster assumption of hyper-ellipsoidal and similar sizes prevented from uncovering the clusters that vary in size or concave shapes. When the amount of input data was large, the algorithms further broke down due to their non-linear time complexity and huge input costs. BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) was proposed to remedy this problem (Zhang, Ramakrishnan, & Linvy, 1996). In pre-clustering phase to reduce input size, dense regions of points were represented by compact summaries, and using the centroids of summaries as cluster seeds, each data point was assigned to the summarized cluster with the closest seed in labeling phase. Only using the centroid to redistribute the data, if the clusters did not have uniform sizes and shapes, a number of points in the bigger cluster were labeled as belonging to the smaller cluster since they were closer to the centroid of the smaller cluster. To alleviate their shortcomings, CURE (Clustering Using Representatives) was presented to adopt a middle ground between the all data extremes and the centroid-based approaches. Drawing a random sample from the database to shrink the representative points toward the centroid, CURE could group the clusters of arbitrary shapes and sizes, even avoid some problems associated with outliers (Guha, Rastogi, & Shim, 1998). Ignoring the information about the aggregate interconnectivity of items in two clusters, it failed to account for special characteristics of individual clusters. Accounting for both interconnectivity and closeness in identifying the most similar pair of clusters, CHAMELEON was given to yields accurate results for highly variable clusters using dynamic modeling (George, Han, & Kumar, 1999). However, CHAMELEON was not suitable
Related Content

A Methodology Supporting the Design and Evaluating the Final Quality of Data Warehouses
[www.igi-global.com/article/methodology-supporting-design-evaluating-final/1811?camid=4v1a](www.igi-global.com/article/methodology-supporting-design-evaluating-final/1811?camid=4v1a)

Improved Approximation Algorithm for Maximal Information Coefficient
Shuliang Wang, Yiping Zhao, Yue Shu and Wenzhong Shi (2017). *International Journal of Data Warehousing and Mining* (pp. 76-93).
[www.igi-global.com/article/improved-approximation-algorithm-for-maximal-information-coefficient/173707?camid=4v1a](www.igi-global.com/article/improved-approximation-algorithm-for-maximal-information-coefficient/173707?camid=4v1a)

Combination Forecasts Based on Markov Chain Monte Carlo Estimation of the Mode
[www.igi-global.com/chapter/combination-forecasts-based-markov-chain/25768?camid=4v1a](www.igi-global.com/chapter/combination-forecasts-based-markov-chain/25768?camid=4v1a)

A Query Language for Mobility Data Mining
[www.igi-global.com/article/query-language-mobility-data-mining/49639?camid=4v1a](www.igi-global.com/article/query-language-mobility-data-mining/49639?camid=4v1a)