HASTA: A Hierarchical-Grid Clustering Algorithm with Data Field

Shuliang Wang, School of Software, Beijing Institute of Technology, Haidian, Beijing, China
Yasen Chen, Tencent Inc., Nanshan, Shenzhen, China

ABSTRACT

In this paper, a novel clustering algorithm, HASTA (HierArchical-grid cluStering based on daTA field), is proposed to model the dataset as a data field by assigning all the data objects into quantized grids. Clustering centers of HASTA are defined to locate where the maximum value of local potential is. Edges of cluster in HASTA are identified by analyzing the first-order partial derivative of potential value, thus the full size of arbitrary shaped clusters can be detected. The experimented case demonstrates that HASTA performs effectively upon different datasets and can find out clusters of arbitrary shapes in noisy circumstance. Besides those, HASTA does not force users to preset the exact amount of clusters inside dataset. Furthermore, HASTA is insensitive to the order of data input. The time complexity of HASTA achieves O(n). Those advantages will potentially benefit the mining of big data.

Keywords: Clustering Algorithms, Data Field, Data Mining, HASTA, Potential Value

INTRODUCTION

With the unprecedented growth in the volume, variety and velocity of data collections (United Nations Global Pulse, 2012; McKinsey Global Institute, 2011; Executive Office of the President, 2012), the huge amount of spatial data, collected from radar, satellite, Internet, and so on, has made it costly, sometimes even unrealistic, for users to explore manually (Reshef et al., 2011; Santer, Wigley, & Taylor, 2011). The lack of capability to handle these massive spatial data leads to the prosperity for the development of spatial data mining. Spatial data mining aims to automate the process of understanding spatial data by representing the data in a concise manner and reorganizing spatial databases to accommodate data semantics (Sheikholeslami, Chatterjee, & Zhang, 1998). As one of most important branches, spatial clustering helps researchers explore spatial datasets and discover those densely populated regions, thus motivates the discovery of hidden knowledge. Clustering algorithms group a set of data in a way that maximizes the similarity within clusters and also minimizes the similarity between two different clusters (Karypis, Han, & Kumar, 1999). Different clustering algo-
Algorithms have been widely used in many fields (Kwok, Smith, Lozano, & Taniar, 2002), such as information retrieval, database management and geographic information system.

Being applied in various practical applications, data mining algorithms face lots of challenges in which effectiveness is one of the most critical ones when manage massive spatial data. An effective algorithm should be able to handle noise properly, detect clusters of arbitrary shapes and perform stably and independently from relying too much on both users’ experience and the order of data input (Fränti, Virmajoki, 2006). Clustering efficiency is another significant consideration, especially for large datasets, but it is really unfortunate to find that most of existed algorithms could only satisfy some parts of requirements. Effective algorithms, for example, DENCLUE (Hinneburg, & Keim, 2003), possess high time complexity, while efficient algorithms perform sensitively to noise or rely too much on parameters, such as WaveCluster (Sheikholeslami, Chatterjee, & Zhang, 1998) and Chameleon (Karypis, Han, & Kumar, 1999).

In this paper, a new hierarchical-grid clustering algorithm, HASTA, is proposed on the basis of data field which is given to express the power of an item in the discourse of universe by means of potential function as the physical field does (Wang, Gan, Li., & Li, 2011). In data field, mutual effects among data items are indicated by a field strength function, which might take different forms, such as nuclear form and gravitational form. At any position inside data field, effects from different sources mutually overlay, and the superposed result is named as potential value. Given the same parameters, the potential values in densely distributed regions are much higher than those in sparse regions. With the definition of an appropriate field strength function, the potential value could well reflect the data distribution. By analyzing the changing regularity of potential value, the distribution of data objects is thus explored.

The rest of this paper is going to be organized as follows. Related work is introduced in Section 2. Then, the detailed information about data field is presented in Section 3. In section 4, the motivation and main steps of HASTA are listed. On the performance evaluation of effectiveness and efficiency, HASTA is compared with other existed algorithms in Section 5. At last, conclusion and further research are discussed.

RELATED WORK

Existed clustering algorithms are roughly categorized into four main groups in this paper including distance-based algorithms, density-based algorithms, grid-based algorithms, and model-based algorithms.

Distance-Based Algorithms

Distance-based algorithms depict the similarity between one data object and another by using their distance. The algorithms, such as CLARANS (Ng, Han, 2002), start with an initial partition and use an iteratively optimizing strategy to seek for the optimal clustering solution. Others may not need an initial partition. However, they need to define a unique data structure to store clustering information. Chameleon (Karypis, Han, & Kumar, 1999), for example, uses \( k \)-nearest graph to model the similarity between two data objects, while BIRCH (Zhang, Ramakrishnan, & Livny, 1996) uses the CF-tree instead.

Density-Based Algorithm

Density-based algorithms detect clusters by calculating the density of data objects nearby. Since that the density of data objects makes no assumptions on the shape of clusters, density-based algorithms are able to find out the clusters of arbitrary shape. DBSCAN (Ester, Kriegel, Sander, & Xu, 1996), for example, detects clusters by counting the amount of data objects inside an area of a given distance. DENCLUE (Hinneburg, & Keim, 2003) models the overall density as the sum of influential functions of data objects, and then use a hill-climbing algorithm to detect clusters inside dataset.
Related Content

A Framework for Mining and Querying Summarized XML Data through Tree-Based Association Rules
Mirjana Mazuran, Elisa Quintarelli, Angelo Rauseo and Letizia Tanca (2012). XML Data Mining: Models, Methods, and Applications  (pp. 264-287).
www.igi-global.com/chapter/framework-mining-querying-summarized-xml/60913?camid=4v1a

A Survey on Temporal Data Warehousing
Matteo Golfarelli and Stefano Rizzi (2009). International Journal of Data Warehousing and Mining  (pp. 1-17).
www.igi-global.com/article/survey-temporal-data-warehousing/1820?camid=4v1a

Object-Oriented Features in Oracle
Johanna Wenny Rahayu, David Tanier and Eric Pardede (2006). Object-Oriented Oracle  (pp. 31-50).
www.igi-global.com/chapter/object-oriented-features-oracle/27337?camid=4v1a

Novel Efficient Classifiers Based on Data Cube
www.igi-global.com/article/novel-efficient-classifiers-based-data/1754?camid=4v1a