Information Sciences 345 (2016) 271–293

Contents lists available at ScienceDirect

Information Sciences

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

A fast density-based data stream clustering algorithm with cluster centers self-determined for mixed data

Jin-Yin Chen*, Hui-Hao He

College of Information Engineering, Zhejiang University of Technology, Hangzhou, China

ARTICLE INFO

Article history:
Received 20 March 2015
Revised 18 November 2015
Accepted 27 January 2016
Available online 3 February 2016

Keywords:
Data mining
Mixed attributes
Data stream clustering
Peak field intensity
Mixed distance measure metrics

ABSTRACT

Most data streams encountered in real life are data objects with mixed numerical and categorical attributes. Currently most data stream algorithms have shortcomings including low clustering quality, difficulties in determining cluster centers, poor ability for dealing with outliers' issue. A fast density-based data stream clustering algorithm with cluster centers automatically determined in the initialization stage is proposed. Based on data attribute relationships analysis, mixed data sets are filed into three types whose corresponding distance measure metrics are designed. Based on field intensity-distance distribution graph for each data object, linear regression model and residuals analysis are used to find the outliers of the graph, enabling cluster centers automatic determination. After the cluster centers are found, all data objects can be clustered according to their distance with centers. The data stream clustering algorithm adopts an online/offline two-stage processing framework, and a new micro cluster characteristic vector to maintain the arriving data objects dynamically. Micro clusters decay function and deletion mechanism of micro clusters are used to maintain the micro clusters, which reflects the data stream evolution process accurately. Finally, the performances of the proposed algorithm are testified by a series of experiments on real-world mixed data sets in comparison with several outstanding clustering algorithms in terms of the clustering purity, efficiency and time complexity.

1. Introduction

As one of the most important techniques in data mining, clustering is to partition a set of unlabeled objects into clusters where the objects fall into the same cluster have more similarities than others [18]. Clustering algorithms have been developed and applied to various fields including text analysis, customer segmentation, gene engineering, etc. They are also useful in our daily life since massive data with mixed attributes are now emerging. Typically these data contain both numeric and categorical attributes [11]. For example, the applicant for a credit card involves data of age (integers), income (float), marital status (categorical), etc., forming a typical example of data with mixed attributes. Up to now most researches on data clustering have been focusing on either numeric or categorical data instead of both. Examples include BIRCH [24], *k*-modes [16], fuzzy *K*-modes [17], BFCM [21], TCGA [14], fuzzy *k*-modes [5] and *k*-means based method [22]. Those methods as [4,8,12,13,15,19,20,23,27,30] face problems when clustering data stream with mixed attributes while the data in stream is emerging very quickly.

* Corresponding author. Tel.: +086 13666611145.
E-mail address: chenjinyin@zjut.edu.cn, chenjinyin@163.com (J.-Y. Chen).

http://dx.doi.org/10.1016/j.ins.2016.01.071
0020-0255/© 2016 Elsevier Inc. All rights reserved.

Distance measure metric for numerical values only cannot capture the distance among data with mixed attributes. Also, the representative of a cluster with numerical values is often defined as the mean of the cluster, which however is not possible for other attributes. To deal with this problem algorithms [8,15,20,23,27] have been proposed, most of which are based on partition. These algorithms firstly obtain a set of disjoint clusters and then refine them to minimize a predefined criterion function. The objective is to maximize the intra-cluster connectivity or compactness while minimizing inter-cluster connectivity. However, most partition clustering algorithms are sensitive to the initial number of clusters which is yet difficult to determine without prior knowledge. They are also more suitable for spherical distribution data and cannot handle outliers.

Inspired by the density-based data clustering, we propose a novel self-adaptive peak density clustering algorithm for data with mixed attributes (ACC-FSFDP). An efficient distance evaluation method is designed based on data types determined in prior. Mixed data sets are filed into three types including numeric dominant data, categorical dominant data and balanced data. Then corresponding distance measure metrics are designed. Based on theory analysis on data field intensity of cluster center and data objects, filed intensity peaks represent the cluster center that always have larger distance from objects with higher field intensity. After the cluster centers have been found, each remaining data object is assigned to the same cluster as its nearest neighbor of higher field intensity, and then final clustering result would be precisely concluded. We also proposed an algorithm extends ACC-FSFDP to data streaming, called Str-FSFDP.

The main contributions of this work are:

1. A novel distance metric for mixed data stream is designed according to data attributes relationships. The ACC-FSFDP algorithm is adopted for initialization, in which lustering centers are determined automatically through two steps: (a) regression analysis techniques is applied for fitting the relationship of field intensity and distance of every data object, (b) residual analysis is used to determine the centers.
2. In Str-FSFDP, a new micro cluster characteristic vector is introduced to maintain the arriving mixed data objects dynamically. The frequency histogram is adopted to record the categorical attributes, and the mean value of numerical attributes and the maximum frequency of categorical attributes are used to present the center of micro clusters.
3. An online/offline framework is adopted for Str-FSFDP, in which the micro clusters decay function and deletion mechanism of micro clusters are applied to maintain the micro cluster in the online stage, which makes Str-FSFDP more consistent with the intrinsic characteristics of the original mixed data stream. The improved density-based method is adopted, in which the micro clusters are considered as a virtual object and cluster them to get final clustering results.

The rest of this paper is organized as follows. Section 2 introduces background information on data clustering stream clustering. Section 3 depicts ACC-FSFDP and Str-FSFDP in detail. Section 4 presents the simulations and analysis of ACC-FSFDP and Str-FSFDP with other outstanding data clustering algorithms. Finally Section 5 concludes the paper.

2. Related works

In order to cluster data with mixed attributes, Huang proposed *k*-prototypes [15] which combine *k*-means and *k*-mode algorithms. Considering the uncertainly character of data, KL-FCM-GM [23] extends *k*-prototypes algorithm. KL-FCM-GM is an extension of Gath-Geva, which is designed for the Guss-Multinomial distributed data. EKP [30] is developed by introduced based on an evolutionary algorithm framework to help *k*-prototypes improve global search capability. Distance-based Agglomerative Clustering algorithm (SBAC) [8] was proposed adopting the distance measure defined by Goodall. CAVE [11] is designed for clustering mixed data based on the variance and entropy. However, CAVE needs to build the distance hierarchy for each categorical attribute, while the determination of distance hierarchy requires the domain expertise. Another *k*-means type algorithm [4] is implemented to deal with mixed data by using co-occurrence of categorical values to calculate the significance of attribute and the distance between categorical values. IWKM [19] is presented which combines mean value of all distribution centroids to represent the prototypes of the cluster and takes into account the significance of each attribute towards the clustering process. WFK-prototypes [20] combine mean value of fuzzy centroids to represent the prototypes of the cluster and adopt the significance concepts proposed by paper [4] to extend *k*-prototypes. SpectralCAT [13] is brought up for clustering numerical and nominal data. Paper [27] proposed a mixed data clustering algorithm based on a unified distance metric without knowing cluster number in prior, and the embedded competition and penalization mechanisms are used to determine the number of clusters automatically by gradually eliminating the redundant clusters. Other clustering algorithms [26] introduced combine the field theory and traditional clustering algorithm to achieve clustering. A hierarchical clustering method [26] based on data fields, which adopts the potential function to describe effect relationship between data objects. FTSC [27] is a field-theory based spatial clustering method, in which a novel concept of aggregation force is utilized to measure the degree of aggregation among the data objects. Alex and Alessandro [1] proposed a new approach based on the idea that cluster centers are characterized by a higher density than their neighbors and by relatively large distance from objects with higher densities. But their approach requires human supervision to determine the cluster centers, and the clustering quality is sensitive to the parameter cutoff distance, in addition, the algorithm cannot deal with the mixed data.

Also a large number of approaches have been proposed for data steam clustering. CluStream [2] is designed for clustering evolving data stream through online/offline two-stage framework for the first time. Micro-cluster structure is defined in the online stage, which maintains the arriving data objects constantly and generates online summary information, while the offline stage is responsible for answering users' requests. The CluStream gets much attention because of its flexibility and