Knowledge-Based Systems 000 (2016) 1–19

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Fat node leading tree for data stream clustering with density peaks

Ji Xu[a,c,d], Guoyin Wang[b,*], Tianrui Li[a], Weihui Deng[c], Guanglei Gou[a]

[a] School of Information Science & Technology, Southwest Jiaotong University, Chengdu 610031, China
[b] Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[c] Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences,Chongqing 400714, China
[d] School of Information Engineering, Guizhou University of Engineering Science, Bijie 551700, China

ARTICLE INFO

Article history:
Received 23 June 2016
Revised 20 December 2016
Accepted 28 December 2016
Available online xxx

Keywords:
Data stream clustering
Density peaks
Fat node leading tree
Change point

ABSTRACT

Detecting clusters of arbitrary shape and constantly delivering the results for newly arrived items are two critical challenges in the study of data stream clustering. However, the existing clustering methods could not deal with these two problems simultaneously. In this paper, we employ the density peaks based clustering (DPClust) algorithm to construct a leading tree (LT) and further transform it into a fat node leading tree (FNLT) in a granular computing way. FNLT is a novel interpretable synopsis of the current state of data stream for clustering. New incoming data is blended into the evolving FNLT structure quickly, and thus the clustering result of the incoming data can be delivered on the fly. During the interval between the delivery of the clustering results and the arrival of new data, the FNLT with blended data is granulated as a new FNLT with a constant number of fat nodes. The FNLT of the current data stream is maintained in a real-time fashion by the Blending-Granulating-Fading mechanism. At the same time, the change points are detected using the partial order relation between each pair of the cluster centers and the martingale theory. Compared to several state-of-the-art clustering methods, the presented model shows promising accuracy and efficiency.

1. Introduction

Data streams have been generated everywhere nowadays because of the technological development in sensors, networks, smart phones and surveillance. Mining data streams in specific domains such as environmental monitoring, city traffic load monitoring [1], or online commercial activities [2], etc., has produced a lot of researches. Clustering data stream has become one of the important issues since a majority of the data streams come unlabeled in the age of Big Data, and turned to be critical in summarizing data or finding out outliers [3].

The major challenges in clustering data streams include: 1) Data streams continuously flow in, so it is usually unfeasible to store all the original data on disk. Therefore, it demands that the data be processed in one single pass. 2) The patterns may change occasionally or frequently as data points in the streams keep arriving [4]. To address these challenges, quite a few research works have been published, e.g. [5–12]. We will discuss these works in the coming section.

Current data stream clustering approaches fall into two categories: *K*-means-like and density- based. The former intends to minimize the distance summation of non-center data points to their corresponding centers, hence the incapability of detecting non-spherical clusters. The methods of latter category cluster items based on their density distribution in the space where the items are embedded, so they can detect right clusters in arbitrary shapes of datasets. However, some density-based data stream clustering methods (like D-Stream [8] and MR-Stream [7]) that find clusters with the concept of *dense grids* (determined with a preset threshold value), may fail to perform well when there coexist clusters of different density levels. Recently, Hahsler and Bolanos addressed this problem by proposing a micro-cluster-based data stream clustering method that leverages the density between micro-clusters through a shared density graph (this method is named as DBSTREAM) [11].

We present in this paper a novel data stream clustering (named as DP-Stream) with the underlying *leading tree* (LT, refer to Section 3.2 for a detailed explanation) structure in the density-peaks-based clustering method [13]. The initial buffered data points are firstly used to construct an LT. The LT can be used

* Corresponding author.
E-mail addresses: alanxuch@hotmail.com, AlanxuCh@hotmail.com (J. Xu), wanggy@ieee.org (G. Wang), trli@swjtu.edu.cn (T. Li), dengweihui@cigit.ac.cn (W. Deng).

http://dx.doi.org/10.1016/j.knosys.2016.12.025
0950-7051/

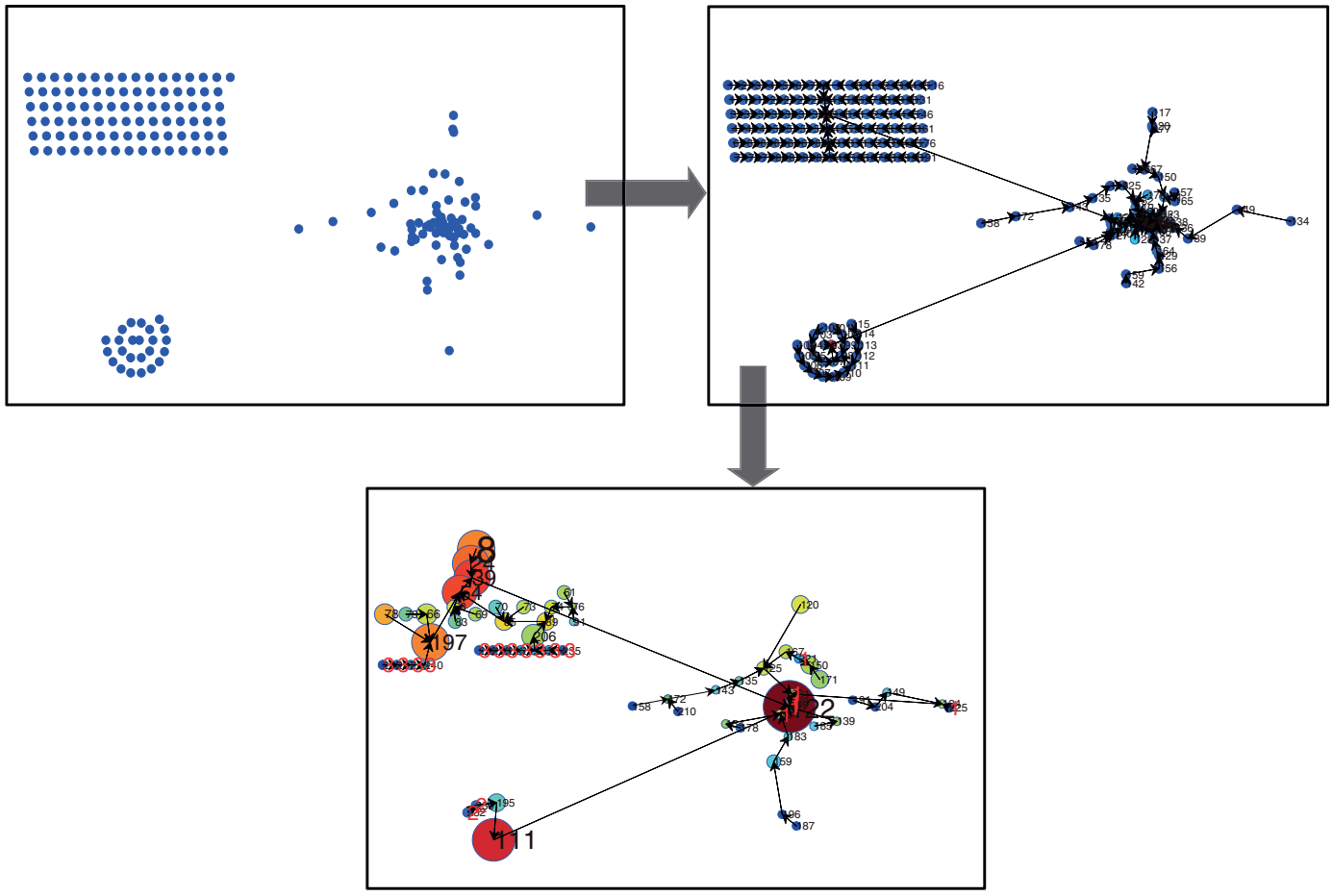

Fig. 1. A simplified illustration of DP-Stream. From top to bottom: buffered data before initialization → initial LT → Granulating LT into FNLT and incorporating the new items with cluster label immediately delivered. The color represents the local density and the radius represents the population (weight) of each node.

to deliver the clustering result for the initial buffered data given the cluster centers are selected automatically. Then the LT is granulated into a *fat node leading tree* (FNLT, refer to Section 4.1 for a detailed definition) by merging the closest points to their corresponding parents, so as to capture the essence of the finest-grained data items with a synopsis of data [14]. Heinz and Seeger proposed to use a Cluster Kernel to present a group of objects in the data stream [15]; they also addressed the issue of limiting the memory consumption in clustering streaming data, in which the cluster kernels may be regarded as information granules as the fat nodes in our DP-Stream. But the difference is obvious, since the fat nodes in our method are some closely located data points other than the resultant clusters. As the data items streaming in, their clustering assignation is quickly determined as soon as the local density of every node (including new items and the existing fat nodes) is incrementally updated. An example of FNLT is shown in Fig. 1.

At the last stage of a clustering-new-items round, the previous FNLT along with the incorporated new items is granulated again (to keep the population of the nodes stable), waiting the next batch of coming items in the stream. Like most of the data stream clustering, DP-Stream includes a fading out mechanism to focus on recent data points and a change point detection utility to deal with concept drift. However, the difference is that our method has very simple implementation of these two utilities due to the properties of an FNLT structure.

DP-Stream has the following salient features:

- It can detect clusters of arbitrary shapes and different density levels;
- The concept drift is accurately and efficiently detected in a simple way;
- It offers an intuitively interpretable visualization of the evolving synopsis of the data stream;
- The evolution of the FNLT is implemented with an efficient incremental update, thus DP-Stream permanently offers clustering result for streaming in items.

Most of the existing data stream clustering methods, such as DBSTREAM, MR-Stream, CluStream, fall in the online-offline category. However, those online-offline models do not adapt well to some applications (e.g. system monitoring), in which the clusters information is required to be always ready.

To the best of our knowledge, DP-Stream is the first model using a density-peaks-based LT structure to cluster data stream. And more importantly, it is the first data stream clustering method that simultaneously meets the two demands: detecting clusters of any shape and running without an offline component.

The remainder of this paper is organized as follows. After a brief discussion of the related works in Section 2, we present in Section 3 the automatic selection of centers and the leading tree structure with density-peaks-based clustering. In Section 4, we describe the DP-Stream method including the components of outliers' recognition, drift detection, and fading function. In Section 5, we discuss the computational complexity of maintaining the FNLT. Section 6 describes detailed experiments with synthetic and real datasets. A conclusion is given in Section 7.

Please cite this article as: J. Xu et al., Fat node leading tree for data stream clustering with density peaks, Knowledge-Based Systems (2016), http://dx.doi.org/10.1016/j.knosys.2016.12.025