International Journal of Synthetic Emotions
Volume 11 • Issue 2 • January-June 2020

Density-Based Clustering Method for Trends Analysis Using Evolving Data Stream

Umesh Kokate, SP Pune University, India
https://orcid.org/0000-0001-6318-7253
Arviand V. Deshpande, SKN College of Engineering, India
Parikshit N. Mahalle, SKN College of Engineering, India

ABSTRACT

Evolution of data in the data stream environment generates patterns at different time instances. The cluster formation changes with respect to time because of the behaviour and members of clusters. Data stream clustering (DSC) allows us to investigate the changes of the group behaviour. These changes in the behaviour of the group members over time lead to formation of new clusters and may make old clusters extinct. Also, these extinct old clusters may recur over time. The problem is to identify and record these change patterns of evolving data streams. The knowledge obtained from these change patterns is then used for trends analysis over evolving data streams. In order to address this flexible clustering requirement, density-based clustering method is proposed to dynamically cluster evolving data streams. The decay factor identifies formation of new clusters and diminishing of older clusters on arrival of data points. This indicates trends in evolving data streams.

KEYWORDS

Big Data, Data Stream Clustering, Density-Based Clustering, Trends Analysis

INTRODUCTION

Nowadays huge data is generated across the various domains in real time, which is high-dimension in nature. Multi-dimensional data streams are generated by most of the applications deployed for whether monitoring, stock trading, telecommunication, network intrusion detection, remotely sense data of planets, and tools for analysis of web. The data streams have temporal order and can only be scan only once (Guha, S. et al., 1998; Yang, J., 2003). There has been active research regarding storage, query and analysis of evolving data streams.

Clustering is one of the major tasks in data mining. Data Stream clustering which is ordered sequence with respect to time-stamped data points in multi-dimension is considered. Data stream clustering has more issues and challenges as compared to traditional data clustering. The challenges are like; data can be scanned and examined in only one pass as data arrive in streams. In many applications, it is essential to know evolving nature of data rather than representing clusters for whole data stream. In most of the cases, data streams were considered as continuous model of static data and implemented clustering algorithms using single-phase (Stonebraker, M. et al, 1993). Such algorithms divides the whole set of data stream into batches and most of them uses k-means clustering

DOI: 10.4018/IJSE.2020070102

algorithms in this finite batch of data (Guha, S. and Mishra, N., 2016; O'callaghan, L. et al., 2002). These algorithms were not in a position to identify the evolving characteristics of data stream. Some of the algorithms try to solve this issue by deploying moving window technique. This again gives partial results in most of the cases (Guha, S. and Mishra, N., 2016; O'callaghan, L. et al., 2002).

Data stream clustering methods proposed by (Aggarwal, C.C. et al., 2004) implemented data stream clustering using two-phase methods, online and offline methods. During online phase data stream is quickly processed and statistical summary is calculated and then during offline phase the same summary is used to generate clusters. The methodology and procedures regarding division of time horizon and statistics management are implemented. This is shown in CluStream (Guha, S. et al., 1998). Most of the data stream algorithms are using two-phase approach similar to CluStream. Semi-Partitioning method is deployed for improved offline phase by (Wang, Z. et al., 2004). Clustering of set of data streams as well as distributed data streams as an extension of work is also mentioned. As CluStream and related algorithms uses k-means method during offline phase, there are number of limitations such as, k-means identify only spherical clusters and not able to detect arbitrary shape clusters, k-means algorithm may not able to detect noise or outliers effectively, it requires number of scans of data, and thus it is not possible to apply directly to large volume of data stream. In CluStream algorithm online phase processes raw data to generate micro-clusters, and these clusters are then used as basic elements during offline phase for further refinement of clusters.

Clustering of data stream using density-based strategy has been widely used and another major methodology in clustering algorithms. In density-based clustering it is possible to identify arbitrary shaped clusters, it can remove noise or outliers and it is possible to scan data only once in order to examine raw data. This method is natural and referred as basic clustering technique for data stream clustering application. As compared to k-means methods density-based clustering does not require prior knowledge of number of probable clusters. DenStream (Cao, F. et al., 2006) algorithm was proposed which calculate density of each data points, and based of certain threshold values the data points are grouped to form a cluster. This requires two phases to implement the clusters. During First Phase, on-line computations are carried out in orders to gather statistical information, this step should be quick and fast as evolving nature of the data stream does not allow to retain the data records for much more time, thus micro-clusters are formed. During Second phase, off-line processing is performed on micro-clusters in order to generate macro-clusters, this leads to formation of arbitrary shape clusters.

In this research work, we propose algorithms to identify trends in evolving data streams which uses D-Stream algorithm (Chen, Y. and Tu, L., 2007), which is a density grid-based clustering framework for data streams. In k-means algorithm, data stream is considered as long sequence of static data set, but the main interest lies in identifying evolving patterns or trends in case of temporal feature of the data stream. The concept of decay factor with respect to the density of data points is introduced for detecting dynamic nature of clusters.

In case of ClusStream architecture it is necessary to explicitly mention the time-duration for clustering, whereas in case of D-Stream algorithm, decay factor of the density which is associated with each data point automatically identify dynamically evolving clusters. This is achieved by calculating weights on the most recent data while considering historical information. Prior information regarding number of clusters is not essential for D-Stream algorithm; therefore it works with very less domain knowledge of application data. In case of evolving data stream, volume of the data is very large; it is not possible to record information related to density for each data. It is therefore, proposed to use grids i.e. set of small cells, to map data records into corresponding grid. The advantage of using grids to map the data records is, storage of raw data is not essential whereas only information regarding new data which was mapped to corresponding grid need to be maintained. In case of multi-dimensional data, there is substantial increase in number of grids. The issue is to handle high dimensionality and scalability of such data. It is observed that, in most of the cases, majority of the grids are empty of contains fewer data records. In D-Stream algorithm these issues are addressed by in depth study of relationship between data density, decay factor and time horizon. There is a unique method that exists