Advanced Science and Technology Letters
Vol.83 (ISA 2015), pp.30-35
http://dx.doi.org/10.14257/astl.2015.83.07

# Towards New Heterogeneous Data Stream Clustering based on Density

Chen Jin-yin, He Hui-hao

Zhejiang University of Technology, Hangzhou,310000
chenjinyin@zjut.edu.cn

**Abstract.** Heterogeneous data stream clustering is an important issue in data stream mining, for the accuracy of the existing heterogeneous clustering algorithm is not high, and don't have a common distance measure method, a heterogeneous data stream clustering algorithm based on the density with mixed distance measure method is proposed. HDSDen algorithm adopts an online/offline two-stage processing framework. According to the situation of dominant property, the online stage use corresponding distance measure method to define the core points among the arriving points, the purpose of the different distance calculation method is to reduce the influence of the non-dominant property on the whole clustering accuracy.

**Keywords:** Data stream; mixed attributes; data clustering; density

## 1 Introduction

With the continuous development of communication technology and hardware equipment, in many emerging field, such as real-time monitoring system, meteorological satellite remote sensing, network traffic monitoring, etc., continuously produce large amounts of data all the time. Those data is different with the traditional data, they are massive, timing, and changing rapidly stream data, and most of the data in the real world is heterogeneous, which include continuous attributes and categorical attributes. Continuous attribute data is the value of the attribute is a continuous, such as length, temperature, etc. Categorical attribute data refers to the value of the property for a limited state, such as color, occupation, etc. Traditional clustering algorithm can’t deal with the data stream, data stream clustering algorithm proposed new requirements are as follows [1]: 1. It has no assumption on the number of clusters; 2. It can discover clusters with arbitrary shapes. 3. It has the ability to handle outliers. Therefore, clustering the data stream has been widespread concerned, and how to analysis and mining valuable information from heterogeneous data stream is becoming more and more important.

In recent years, a lot of data clustering algorithms appear, but most of the existing algorithms limited processing the continuous attributes data stream [2-6], in addition, there are few algorithms limited processing the categorical attributes data stream [7], and less algorithms for mixed attributes data stream. Aggarwal, etc. proposed an algorithm framework CluStream [2] for evolving data stream, which adopts two-stage

ISSN: 2287-1233 ASTL

processing framework for the first time: online-micro-clustering and offline-macro clustering. The online stage proposed micro-cluster structure, and maintenance arriving data points constantly, generate summary information. The offline stage responsible for the user request, to produce the final clustering results based on summary data. The flexible scalability of algorithm get the majority of attention. But Clustering algorithm still exist some disadvantages, firstly, it can’t discover cluster with arbitrary shapes, secondly, poor adaptability to noise, finally, it requires people to specify the number of clusters of micro cluster, which impact the shape of the distribution of the original data seriously. For those problems, Aggarwal etc. proposed HPStream algorithm [3] based on the CluStream, the algorithm aim for high dimensional data stream, introduced projection and decay function, which have a better effect on high dimensional clustering anlysis than CluStream. Cao etc. proposed Den-Stream algorithm [4], the algorithm follow CluStream, which use the two-stage framework, and introduced potential c-mirco-cluster and outlier micro-cluster structure, which can discover cluster with arbitrary shapes.

For the problems on the existing algorithms, this paper proposed a density-based heterogeneous data stream clustering algorithm with mixed distance measure method. According to the situation of dominant property, the online stage use corresponding distance measure method to define the core points among the arriving points. All the density-reachable points form a cluster in the offline stage, and put all the not-clustered points into the reservoir, and the number of the reservoir exceeds the threshold value, we will re-cluster the points to improve the accuracy of clustering. Experiments on real data sets show that the algorithm can achieve better clustering results, and give the clustering results at any time, which can deal with the heterogeneous data stream efficiently.

## 2 The traditional density-based clustering and related definitions

The traditional density based clustering algorithm[11] is an algorithm which based on density to search for dense area, the purpose of the algorithm is to find the core points according to the parameters ε (ε-neighborhood) and μ (density threshold), and from the core points put all the density-connected points form a cluster. The algorithm use Euclidean distance to measure the similarity between points, the distance formula as follows:

$$d(X_i, X_j) = \sqrt{\sum_{p=1}^{n}(X_{ip} - X_{jp})^2}$$

(1)

Where n represents the number of dimensions of data.

Related concepts in the algorithm are defined as follows:

**Definition 1(Core point):** A core point is defined as a point, in whose ε-neighborhood the number of points is at least an integer μ.

**Definition 2(Border point):** A border point is defined as a point, which is not a core point, but located in the core points’ ε-neighborhood.

**Definition 3(Noise point):** A noise point is defined as a point, which neither a core point, nor a border point.

**Definition 4(directly density-reachable):** A point p is directly density-reachable form a point q, if point p is in q' ε-neighborhood and q is a core point.

**Definition 5(density-reachable):** A point p is density-reachable from a point q, if there is a chain of points p1, p2 … pn, p1=q, pn=p such pi+1 is directly density-reachable from pi.

**Definition 6(density-connected):** A point p is density-connected to a point q, if there is a point o such that both p and q are density-reachable from o .

In order to define the core point p, algorithm according to formula (1) to calculate the distance between other points q and p one by one, statistics the number whose d(p, q)< ε, and then label the point p as a core point or border point or noise point, until all the points labeled, the algorithm from the core point, put all density-connected points together to form a cluster.

## 3 HDSDen algorithm framework and related concepts

We need to define some symbols which used in this paper, the data stream like a data sets D={X1, X2 … Xi …}, and the arrival time of the points represented as T1, T2… Ti …, every point have d dimensions, include c dimensions continuous attributes and b dimensions categorical attributes, represented as Xi = Ci : Bi = $(x_i^1, x_i^2, \dots, x_i^c, y_i^1, y_i^2, \dots, y_i^b)$.

### 3.1. Algorithm framework

HDSDen composed of two-parts: online-maintenance and offline-cluster. In the online stage, we maintenance the arrival points constantly, update the information of points' ε-neighborhood. In the offline-stage, according to the user request, cluster the online summary information at a time, then give the corresponding clustering results. Co-operation with Online / offline two stages, dynamic and fast processing streaming data, which is good to meet user' demand for data stream analysis. Mining model is described as fig.1.

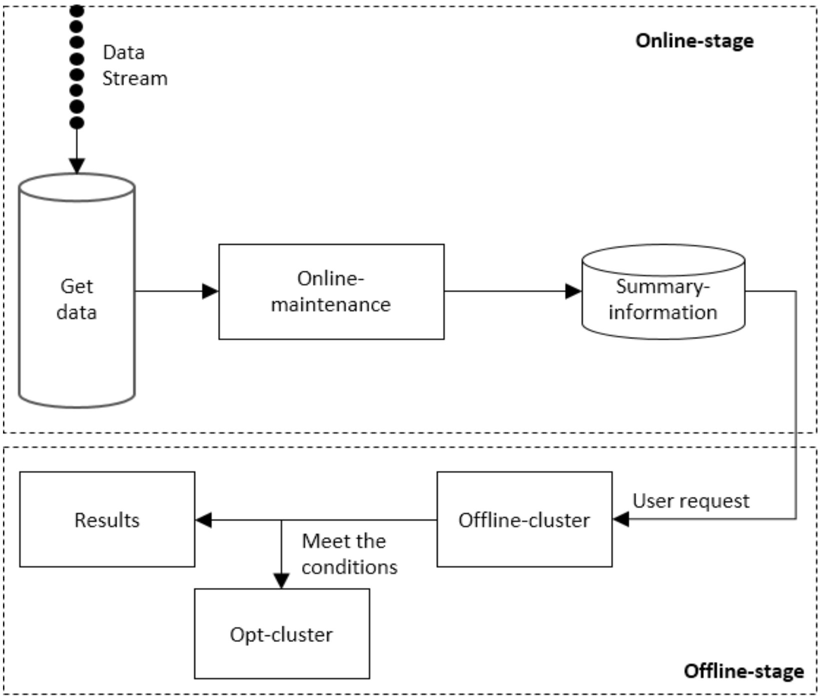

**Fig.1** Data Stream Clustering model of HDSDen

### 3.2. Distance measure method

The traditional density based method is only to deal with the continuous attributes data stream, Euclidean distance can't calculate the mixed attributes which not only include continuous attributes, but also include categorical attributes, therefore, we do some improvement for the distance measure of heterogeneous data.

Consider that heterogeneous data include continuous attributes and categorical attributes, which must exist the situation of continues attributes is the dominant attribute or categorical attributes is the dominant attributes. For this problem, we give two different distance measure methods; the purpose of different distance measure methods is to reduce the influence of the non-dominant attribute on the whole clustering accuracy.

This paper according to the features of heterogeneous data, use d(Xi, Xj)n and d(Xi, Xj)c to represent the distance of continuous part and the distance of categorical part. The definitions as follows (data sets D as an example):

1. If data sets D' continuous attributes are dominant attribute, we define the distance between points as follows:

**Definition 7:** Any two points Xi, Xj the distance of continuous part is:

$$d(X_i, X_j)_n = \sqrt{\sum_{p=1}^{c}(X_{ip} - X_{jp})^2} \quad (2)$$

**Definition 8:** For every dimension of continuous part of any two points Xi, Xj, we adopt dualistic approach, the distance of the pth dimension of Xi, Xj is:

$$d(X_{ip}, X_{jp})_c = \begin{cases} 0 & X_{ip} = X_{jp} \\ 1 & X_{ip} \neq X_{jp} \end{cases} \quad (3)$$

And the distance of categorical part is:

$$d(X_i, X_j)_c = \sum_{p=1}^{b} d(X_{ip}, X_{jp}) \quad (4)$$

2. If data sets D' categorical attributes are dominant attribute, we define the distance between points as follows:

**Definition 9:** For every dimension of categorical part of any two points Xi, we adopt standardized approach, the distance of the pth dimension of Xi is:

$$d(X_{ip})_n = \frac{X_{ip} - X_{ip\_min}}{X_{ip\_max} - X_{ip\_min}} \quad (5)$$

Where ip_max represent the max number of this dimension， ip_min represent the min number of this dimension.

And the distance of continuous part is:

$$d(X_i, X_j)_n = \sum_{p=1}^{c}(d(X_{ip})_n - d(X_{jp})_n) \quad (6)$$

**Definition 10:** For every dimension of continuous part of any two points Xi, Xj, we adopt dualistic approach, the distance of the pth dimension of Xi, Xj is:

$$d(X_{ip}, X_{jp})_c = \begin{cases} 0 & X_{ip} = X_{jp} \\ 1 & X_{ip} \neq X_{jp} \end{cases} \quad (7)$$

And the distance of categorical part is:

$$d(X_i, X_j)_c = \sum_{p=1}^{b} d(X_{ip}, X_{jp}) \quad (8)$$

**Definition 11:** Any two points Xi, Xj in data sets D, we define the distance as:

$$D(X_i, X_j) = d(X_i, X_j)_n + d(X_i, X_j)_c \quad (9)$$

As Xi an example, calculate the distance between the other points in data sets D and point Xi, if the distance lower than ε, we put the point into Xi' ε-neighborhood.

## 4 Conclusion

For the accuracy of the existing heterogeneous clustering algorithm is not high, and don't have a common distance measure method, this paper proposed a new heterogeneous data stream clustering algorithm based on density with mixed distance measure method. This paper proposed re-clusters strategy, which can re-cluster the noise points to improve the clustering accuracy.

## References

1. Zhu, Q., Zhang, Y.-H., Hu, X.-G., Li, P.-P.: A double-window-based classification algorithm for concept drifting data streams [J]. Acta Automatica Sinica, 37(9):1077-1084 (2011)
2. Aggarwal, C. C., Han, J. W., Wang, J. Y., Yu, P. S.: A framework for clustering evolving data streams [J]. In: proceedings of the 29th International Conference on Very Large Data Bases (Vol. 29), VLDB Endowment, 81-92 (2003)
3. Aggarwal, C. C., Han, J. W., Wang, J. Y., Yu, P. S.: A framework for projected clustering of high dimensional data streams [J]. In: proceedings of the 30th International Conference on Very Large Data Bases (Vol. 30), VLDB Endowment, 852-863 (2004)
4. Cao, F., Ester, M., Qian, W., et al.: Density-based clustering over an evolving data stream with noise[C]. Proc of the SIAM Conf on Data Mining. Bethesda, 326-337 (2006)