International Journal of Database Theory and Application
Vol.8, No.3 (2015), pp.167-178
http://dx.doi.org/10.14257/ijdta.2015.8.3.14

Density-Based Heterogeneous Data Stream Clustering Algorithm with Mixed Distance Measure Methods

Chen Jin-yin and He Hui-hao

Zhejiang University of Technology, Hangzhou310000
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. 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.

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 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

ISSN: 2005-4270 IJDTA

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. But because Den-Stream adopts globally consistent parameters, which makes the clustering results are very sensitive to the choice of parameters. Marwan Hassani etc. proposed PreDeConStream algorithm [5], the algorithm also use two-stage processing framework and potential c-mirco-cluster and outlier micro-cluster structure, introduced a factor when clusters update will impact on the adjacent micro-clusters. The algorithm can deal with high dimensional data, but greatly increased time complexity. Zhang etc. proposed StrDenAP algorithm [6] based on StrAP, which adopts two-stage framework and use density-based and affinity propagation techniques. The algorithm can have a good clustering accuracy.

Because of the most of data in the real world are mixed attributes data, for this problem, Yang etc. proposed HCluStream algorithm [8] based on CluStream, defines the histogram description of the categorical attributes in micro-cluster and proposed Poisson Arrival model of data stream. The problem is HCluStream algorithm can't discover cluster with arbitrary shapes. Huang etc. proposed MCStream algorithm [9] based on HCluStream, the algorithm adopts two-stage framework, use oriented dimension distance to measure the similarity between points in the online-stage, in the offline-stage use improved DBSCAN density-based method for the final clustering. The algorithm can discover cluster with arbitrary shapes, but oriented dimension distance needs much specified parameters. Huang etc. proposed DkHDSC algorithm [10], the algorithm use improved K-nearest-neighbors and improved oriented dimension distance, which can get a good clustering result, but still exist the problem which can't 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:

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

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 p_1, p_2 … p_n, p_1=q, p_n=p such p_{i+1} is directly density-reachable from p_i.

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.

4. 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={X_1, X_2 … X_i …}, and the arrival time of the points represented as T_1, T_2… T_i …, every point have d dimensions, include c dimensions continuous attributes and b dimensions categorical attributes, represented as X_i = C_i : B_i = $(x_i^1, x_i^2, \dots, x_i^c, y_i^1, y_i^2, \dots, y_i^b)$.

4.1. 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.

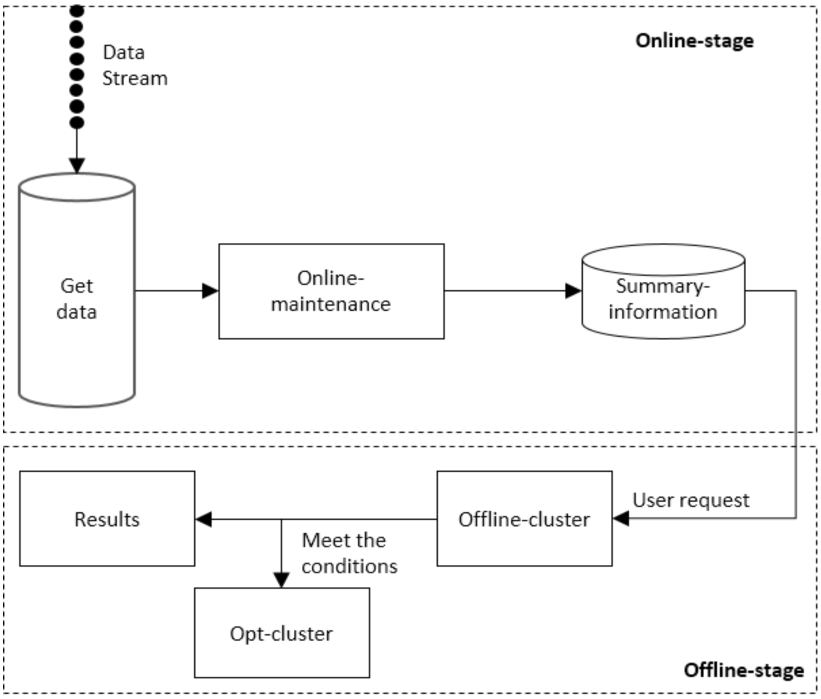

Figure 1. Data Stream Clustering Model of HDSDen

4.2. 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 problems, 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 $d(X_i, X_j)_n$ and $d(X_i, X_j)_c$ 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 X_i, X_j the distance of continuous part is:

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

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

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

And the distance of categorical part is:

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

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 X_i, we adopt standardized approach, the distance of the pth dimension of X_i is:

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

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:

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

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

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

And the distance of categorical part is:

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

Definition 11: Any two points X_i, X_j in data sets D, we define the distance as:

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

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

5. Detailed Description of HDSDen Algorithm

5.1. Pretreatment Process

For the situation of continues attributes is the dominant attribute or categorical attributes is the dominant attributes, we have different distance measure methods, so we should pre-judge the data, and according to the result to change the corresponding attributes. The perform operations show in algorithm 1.

```
Algorithm 1. Pretreatment ()
If (continues attributes is the dominant attribute)
For (arrived every points)
    Change the point information according to section 4.2 and save the result.
End if
If (categorical attributes is the dominant attribute)
For (arrived every points)
    Change the point information according to section 4.2 and save the result.
End if
End
```

The processed result will used in the online/offline stage.

5.2. Online Maintenance

Due to the dynamic characteristics of the data stream, new points appear constantly. When a new point arrival, according to formula (9) calculate the distance between the new point and the arrived points, then we update the ε-neighborhood information of points depend on the distance. The perform operations show in algorithm 2.

```
Algorithm 2. SetApprivalPoints ()
A new point $X_t$
For (every point $X_i$)
    If ($D(X_t, X_i) < ε$)
        Put the $X_i$ into the $X_t$' ε-neighborhood.
        Put the $X_t$ into the $X_i$' ε-neighborhood.
    End if
For (every point $X_i$)
    If (the number of points in $X_i$' ε-neighborhood >= μ)
        $X_i$ is a core point.
    End if
```

```
End
```

The number of core points will become more and more with the time evolve, the neighborhood of core points will cover most points, we need to put all the density-reachable points together to from a cluster in offline stage.

5.3. Offline Stage

Depend on the summary information saved in the online stage, from the core points find all the density-reachable points to form a cluster. Then those non-clustered points will be put into reservoir, when the number of points in the reservoir exceed than threshold value, we re-cluster the points to improve the whole clustering accuracy. The perform operations show in algorithm 3 and algorithm 4.

```
Algorithm 3. Do_cluster ()
Input: core points and their neighborhood information
Output: accurate clustering result at a time
do
    get a untreated point p;
    If (p is a core point) then
        Find all the density-reachable points to form a cluster
    Else
        Break;
    End if
    until all the points treated.
```

If meet the conditions of re-cluster, we will re-cluster the points in the reservoir.

```
    Algorithm 4. Opt_cluster ()
    Input: the points information of reservoir
    Output: accurate clustering result of reservoir points
    For (every points)
        SetApprivalPoints ();
        Do_Cluster();
    End
```

5.4 The Process of HDSDen Algorithm

HDSDen algorithm need two parameters ε and μ to search for core points. If any point p has more than μ points in its ε neighborhood, we create a cluster which regard point p as a core point, then cluster all the directly density-reachable core points, finally find all the density-reachable points from those core points, put them into this cluster. If p is not a core point, we get another untreated point sequentially, until find a whole cluster. Then we find another untreated core point, and find another whole cluster, until all the points treated. Those non-clustered points are noise points at this time, we put them into reservoir, to find another chance to re-cluster them.

Dynamic data stream D, neighborhood ε, density threshold μ, the size of reservoir Max_Num. HDSDen algorithm is described as follows:

1) Pretreatment: use different distance measure methods to handle mixed attributes data, process algorithm 1.

When a new point arrival, update the neighborhood information of points.

2) Online-maintenance: process algorithm 2, SetArrivalPoints().

3) Offline-cluster: process offline cluster algorithm.

If exist non-clustered points, put them into reservoir

4) Opt-cluster: if the number of points in the reservoir than Max_Num, then re-cluster the points.

6. Experimental Evaluation

In this section, we present an experimental evaluation of HDSDen. We implemented HDSDen in Microsoft visual C++. All experiments were conducted on a 2.6GHz Inter Core I5 PC with 4GB memory, running Microsoft windows 7.

6.1. Categorical Attributes Dominant Data

For convenient to compare, experiment 1 use the same data set as paper [10], which is the Zoo data sets of UCI data mining data sets. This data sets include 1 continuous attribute data and 15 categorical attributes, and been divided into 7 classes. We all know that data stream clustering algorithm cannot cluster all the points, which means exist the situation of missing data or clustering quality impure, we use the clustering entropy to evaluate the clustering quality. The results of proposed approach and the comparative methods in paper [10] is shown in fig.2.

The formula of clustering entropy is :

$$\mathrm{Entr}(C_i) = -\frac{1}{\log N}\sum_{t\in T}\frac{N_{ti}}{N_i}\log(\frac{N_{ti}}{N_i}) \qquad (10)$$

N is the number of data set, N_i represents the number of the cluster C_i, N_{ti} represents the number of the main cluster in the cluster C_i. The range of $\mathrm{Entri}(C_i)$ is [0, 1], 1 represents all kinds of cluster is evenly distributed, and 0 represents all kinds of cluster is composed by one cluster, so the lower the result , the better the clustering quality.

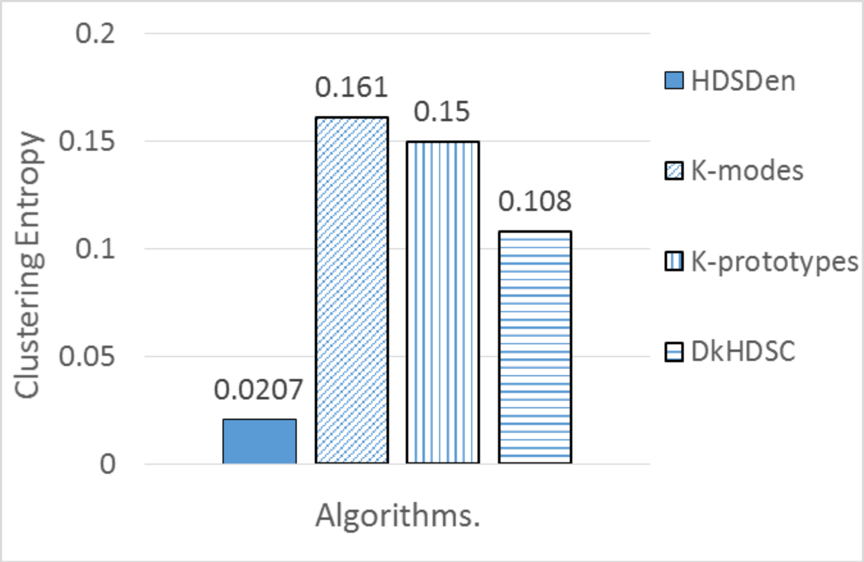

Figure 2. Clustering Entropy Comparison of Methods

The results shown in Fig.2 explains that the clustering quality of proposed approach is better than other algorithms.

And Table.1 is the results of algorithms' time performance, which shows that the proposed approach is a little slower than others, because the proposed approach need to change the dimension information according to the dominant attributes.

Table.1 Execution Time

Algorithms	HDSDen	K-prototypes	D_kHDSC
Time / ms	50	40	31.5

6.2. Continuous Attributes Dominant Data

Experiment 2 use the KDD-CUP 99 Network Intrusion data set of UCI data mining data sets. This data set contains 41 dimensions, which include 34 dimensions continuous attributes and 7 dimensions categorical attributes, and every record has been label into 5 big classes and 24 small classes, include the normal connection and different intrusion and attack.

Experiments test the clustering quality of HDSDen, and compare the performance with CluStream[2], HCluStream[8] and MCStream[9]. Unless particularly mentioned, the parameters of HDSDen adopt the follow setting: ε-neighborhood = 20.3, density threshold μ = 4, speed=200. The parameters for CluStream, HCluStream and MCStream are chosen to be the same as those adopted in [2, 8, 9].

6.2.1. Clustering Quality Evaluation: The clustering quality is evaluated by the average purity of clusters which is defined as follows:

$$\text{Pur} = \sum_{i=1}^{k} \frac{|C_i^d|}{|C_i|} / K \tag{11}$$

Where k denotes the number of clusters. $|C_i^d|$ denotes the number of points with the dominant class label in cluster i. $|C_i|$ denotes the number of points in cluster i. intuitively, the purity measures the purity of the clusters with respects to the true cluster (class) labels that known for our data sets.

We chose to experiment on some representative time stamp, for example, at time 211 there are 23 ”postsweep” attacks, 1 “phf” attack. At time 1857, there are 79 “smurf” attacks, 99 “teardrop” attacks and 22 “pod” attacks. Fig.3 shows the average clustering quality in different speed and different horizon. When speed =200, H=1, it can be seen that HDSDen clearly outperforms than other algorithms and the purity of HDSDen is always than 95%. And we also experiment on speed=100, H=10 (the data described in Table.2), the results show that HDSDen can also get high clustering accuracy with slower speed and bigger horizon.

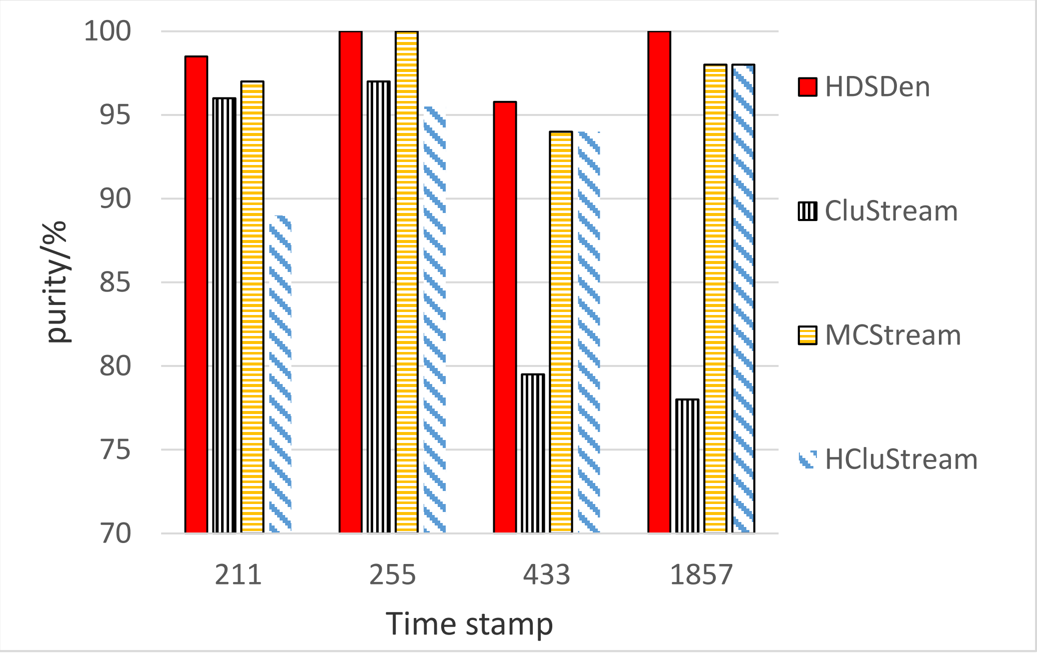

Fig.3 (a) KDD-CUP 99 Data Sets (H=1, v=200)

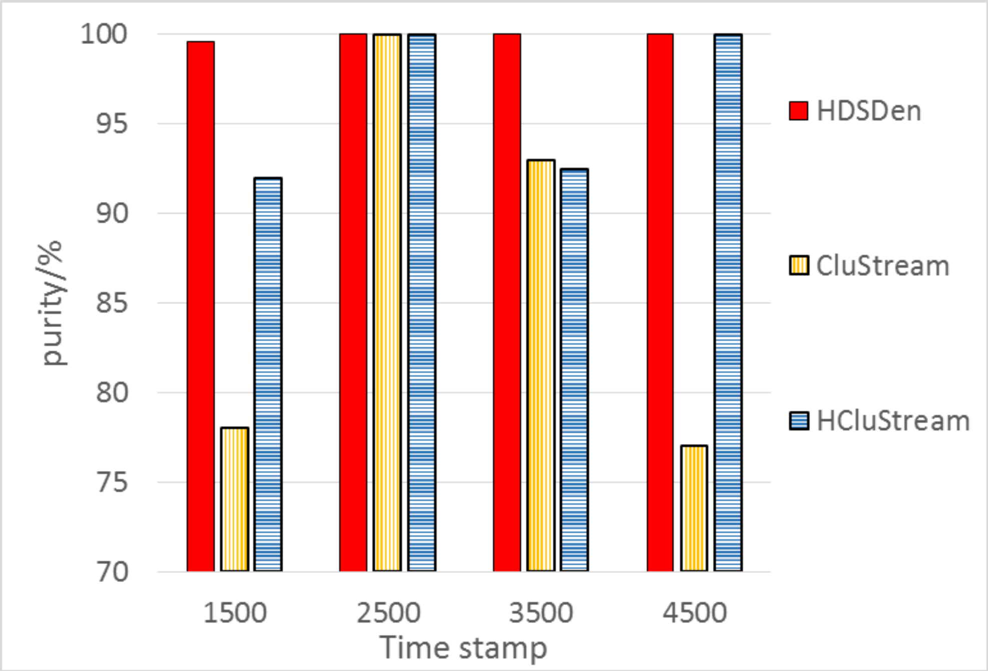

Figure 3. (b) KDD-CUP 99 Data Sets (H=10, v=100)

The high quality clustering results of HDSDen is thanks to the distance measure methods, which can reduce the influence of non-dominant attributes on the whole clustering accuracy, and make the HDSDen better than other algorithms. At the meantime, it also thanks to the re-cluster strategy, which can give the noise points another chance to form a cluster, and improve the quality of the clustering.

Table 2. Number of Categories on Invasion Time

The number of attack	Intrusion time （v=100 h=10）			
	1500	2500	3500	4500
Normal	373		381	215
Satan	380			
Bufoverflow	5			
Teardrop	99			
Smurf	143	1000		785
Neptune			618	
Land			1	
Total	1000	1000	1000	1000

6.2.2. Execution Time: Because HDSDen compared to CluStream adds a process of handling with categorical attributes, and cost much time in pretreatment and re-cluster process, so the calculate speed is lower than CluStream, CluStream can deal with 4000-5000 points per second, under conditions without special optimized, HDSDen can deal with 1000 points per second, and processing 10000 data takes 10 seconds. Fig.4 shows the result.

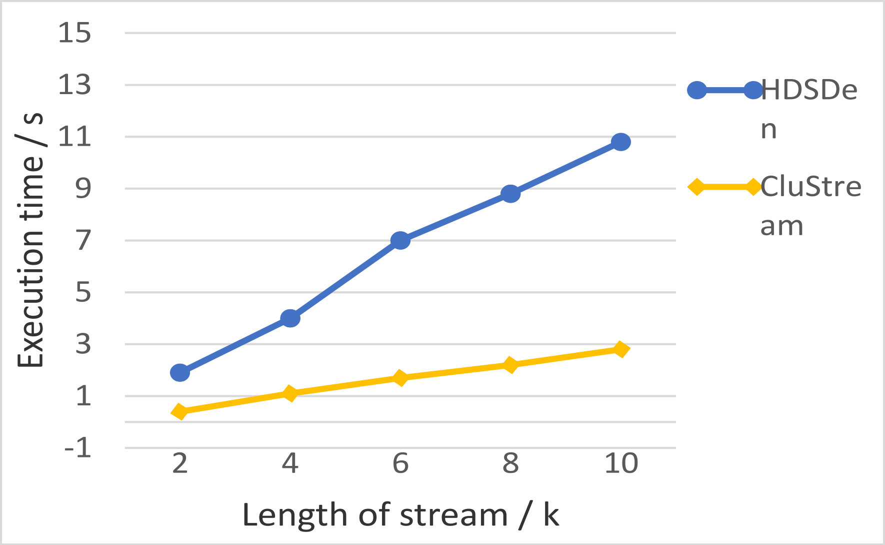

Figure 4. Execution Time Comparison (h=1, v=200)

7. 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, and HDSDen has some features: 1. Consider the mixed attributes data has different situations of dominant attributes, we introduced mixed distance measure method according to the dominant attributes, which can reduce the influence of non-dominant attributes on the whole clustering accuracy, and get a better clustering quality. 2. This paper proposed re-cluster strategy, which can re-cluster the noise points to improve the clustering accuracy. And finally the experiments show that HDSDen has a good clustering quality and can give the clustering results at any time, capable of handling mixed attributes data stream clustering problem effectively.

Acknowledgements

The author would like to thank the anonymous reviewers for their valuable comments and suggestions that helped improve the quality of this paper.

References

[1] Q. Zhu, Y.-H. Zhang, X.-G. Hu, P.-P. Li, “A double-window-based classification algorithm for concept drifting data streams”, Acta Automatica Sinica, vol. 37, no 9, **(2011)**, pp. 1077-1084.

[2] C. C. Aggarwal, J. W. Han, J. Y. Wang, P. S. Yu, “A framework for clustering evolving data streams”, Proceedings of the 29th International Conference on Very Large Data Bases, vol. 29, VLDB Endowment, **(2003)**.

[3] C. C. Aggarwal, J W. Han, J. Y. Wang, P. S. Yu, “A framework for projected clustering of high dimensional data streams”, Proceedings of the 30th International Conference on Very Large Data Bases vol. 30, VLDB Endowment, **(2004)**, pp. 852-863.

[4] F. Cao, M. Ester, W. Qian, “Density-based clustering over an evolving data stream with noise”, Proc of the SIAM Conf on Data Mining. Bethesda, **(2006)**, pp. 326-337.

[5] M. Hassani, P. Spaus, M. M. Gaber, T. Seidl, “Density-based projected clustering of data streams”, Proceeding of the 2012 Scalable Uncertainty management, Berlin Heidelberg, Springer, **(2002)**, pp. 311-324

[6] J.-P. Zhang, F.-C. Chen, S.-M. Li, L.-X. Liu, “Data Stream Clustering Algorithm Based on Density and Affinity Propagation Techniques”, Acta Automatica Sinica, vol. 40, no. 2, **(2014)**, pp. 277-288.

[7] C. C. Aggarwal, P. S. Yu, “A framework for clustering massive text and categorical data streams”, Proceeding of the 6th SIAM International Conference on Data Mining, **(2006)**; Bethesda.

[8] C. Y. Yang, J. Zhou, “A heterogeneous data stream clustering algorithm”, Chinese J of Computers, vol. 30, no. 8, **(2007)**, pp. 1364-1371.

[9] D. C. Huang, T. H. Wu, “Density-based clustering algorithm for mixture data sets”, Control and Decision, vol. 25, no. 3, **(2010)**, pp. 416-421.

[10] D. C. Huang, X. Q. Shen, Y. H. Lu, “Double k-nearest Neighbors of Heterogeneous Data Stream Clustering Algorithm”, Journal of Computer Science and Technology, vol. 40, no. 10, **(2013)**, pp. 226-230.

[11] M. Ester, H-P Kriegel, J. Sander, X Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, Proc. Of KDD, **(1996)**.

Authors

Chen Jin-Yin, was born in 1982, Ph.D., associate professor at Zhejiang university of technology. His research interests cover intelligent computing and its application in power system and network security.

.

HE Hui-Hao, born in 1990, M.S. candidate. His research interests include power system fault diagnosis and data mining.