Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 15 (2011) 1854 – 1858

Procedia Engineering

www.elsevier.com/locate/procedia

A Rough-Set-based Clustering Algorithm for Multi-stream

Haiyan Zhou[a] , Xiaolin Bai [b] Jinsong Shan [a],a*

[a]*Faculty of Computer Engineering, Huaiyin Institute of Technology, Jiangsu, Huaian, 223003,China*
[b] *Department of computer，Taiyuan Normal Univesity，Shanxi,Taiyuan, 030012,China*

Abstract

The paper propose a rough-set-based clustering algorithm for multiple data stream, which solve the problem that existing clustering algorithm for multiple data streams can not take into account conflicts between clustering quality and efficiency. Firstly, the algorithm calculates the distance between data stream to determine the initial equivalence relations, and calculates the similarity between the initial equivalence relation to determine the initial cluster. In the second place, the similarity between the initial clusters is used to merge the initial clusters. Finally, k-means clustering algorithm is called to dynamically adjust the clustering results, and then real-time clustering structure is obtained. In conclusion Experimental results demonstrated that the algorithm has higher efficiency and clustering quality.

Selection and/or peer-review under responsibility of [CEIS 2011]
Keywords:clustering; multiple data stream;rough set

1. Introduction

Cluster analysis and event discovery for data stream is the primary task of studying data stream [1,2]. The current research findings can be used only to process and analysis a single data stream. But in practice, a large number of applications require evolution of data stream clustering analysis among the correlation data streams. The clustering analysis for multiple data streams not only could find the cluster model in local time area, but also could mine the whole event points.

For example, each stock transaction data is a data stream in the stock exchange analysis. Fig.1 shows that the changing curve of three streams of stock data $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$[3], the X-axis presents time, Y-axis for the range of ups and downs. According to the changes in coupling relationship of multiple data streams, the whole time interval is divided into 3 parts by t_1 and t_2. When t <t1, x1 and x2 have consistent changing rate. So in the interval, the clustering model of three data stream is: *{x1, x2}* and *{x3}*. When *t1 <t <t2*, the coupling relationship changed, the rate of change at this time three stocks are more consistent. So clustering model is *{x1, x2, x3}*. When *t> t2*, x1 and x3 have

* Corresponding author. Tel.: +86 517 8359 1046.
E-mail address: zhy_5703@163.com.

doi:10.1016/j.proeng.2011.08.345

more consistent rate of change than before. In the situation, the clustering model is {*x2*} and {x*1, x3*}. The turning point of clustering model t1 and t2 is the event points of evolution.

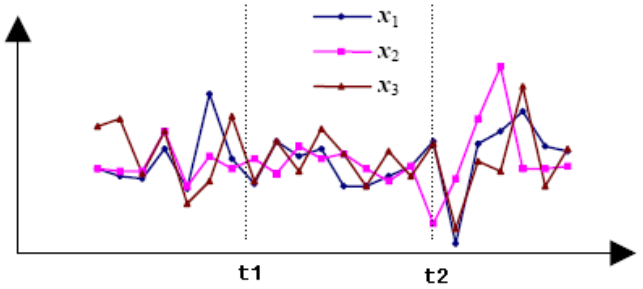

Fig.1 Coupling between streams

To reveal the evolution law and global events, our works are the following: (1) defines the concept of degree of coupling to measure the similarity between the two data streams in terms of data streams changing. (2) Puts forward a rough set based clustering algorithm for multiple data streams RSBCAM. (3) Experiments prove that algorithm RSBCAM is effective and feasible.

2. Related works

Evolution of clustering data streams. Literature [2] was proposed a data stream clustering algorithm based k-Median. Literature [4] further proposed CluStream algorithm, which divides the data stream clustering into online and offline clustering module. The current researches on the data stream have one or more than one of the following shortcomings:

(1) Restrict its use to clustering analysis for a single data stream, and lack of research on clustering of multiple data streams.

(2)Lack of evolution analysis for multiple data streams, in particular, without mining evolution events by clustering models between multiple data streams.

(3) The current results use attribute value to measure similarity of data stream, without considering the history trends of similarity.

In recent years, clustering algorithm based on rough sets has attracted widely research interests, and been successfully applied with cluster high-dimension and sparse data [5].

3. Rough Set Based Clustering model for Multi-stream

3.1. Basic concepts

According to the characteristics of the data stream, the paper only considers time interval composed of discrete time. This article references the concept of time in temporal database , which is defined as a time quantum that is no longer divided into the smallest time interval, its granularity can be second, hour, day and so on.

Definition 1(Changing Rate of data stream).If the average of data stream $\boldsymbol{d}_i$ is $avg(\boldsymbol{d}_i,t)$, then the changing rate of $\boldsymbol{d}_i$ in the time *t is*

$$f(\boldsymbol{d}_i,t)=[avg(\boldsymbol{d}_i,t)-avg(\boldsymbol{d}_i,t-1)]/avg(\boldsymbol{d}_i,t-1). \quad (1)$$

Definition 2(Distance) The distance between the data stream $\boldsymbol{d}_k,\boldsymbol{d}_l$ in the specific period time is

$$d_{kl}(I)=\sqrt{\sum_{i=a}^{b}\{[(f(d_k,i)-F_1]-[f(f(d_l-F_2)]\}^2}\ . \quad (2)$$

In the equation, $F_1=[1/(b-a)]\sum_{i=a}^{b} f(d_k,i), F_2=[1/(b-a)]\sum_{i=a}^{b} f(d_l,i)$. When without ambiguity, $d_{kl}(I)$ can be written in d_{kl}.

The distance reflects the dynamic correlation of the two data streams from a certain historical stage. The dynamic correlation is the similarity of trend. Proved easily that when the value of d_{kl} in $[0,+\infty]$, and $d_{kl}=0$, the data stream d_k, d_l completely positive coupling.

Definition 3(The Similarity of Initial Equivalence Relation). If R_k， R_l are the initial equivalence relation in random, the similarity of them is

$$S(R_k, R_l)=\frac{|[d_k]\cap [d_l]|}{|[d_k]\cup[d_l]|}. \quad (3)$$

Definition 4 (General Equivalence Relation). If R_k， R_l are the initial equivalence relation in random, general Equivalence Relation is $R_k^{'}=\{P_k, D^n-P_k\}$， $P_k=\{d_l \mid S(R_k R_l)\geq\beta_2$ $l=1,2,\ldots,n\}, k=1,2,\ldots,n$. $S(R_k, R_l)$ is the similarity of initial equivalence relation of R_k, R_l， and β_2 is the threshold of similarity of initial equivalence relation.

3.2. Clustering Algorithm

Firstly, the multi-stream algorithm proposed in the paper calculates the distances between each data stream and others. According to the specified threshold, the initial equivalence relation of each data stream calculated to get the similarity of the initial equivalence. By definition 5, the equivalence relation of data stream is modified to get the domain partition, and then merge clusters. Secondly, by calculation results above, dynamic clustering k_means algorithm is called to update the clustering results, which make the data stream clustering results remain the latest state.

Algorithm 1 (Rough Set Based Clustering Algorithm for Multi-stream)

Step 1 (generating Initial Equivalence Relation)

If $D^n=\{d_1,\ldots,d_n\}$ represents data streams set, the multi-stream algorithm calculates the distances $d_{kl}(I)$between each data stream $d_k \in D^n$ and others in D^n. By specified threshold β_1, calculating the initial equivalence relation R_k, $k=1,2,\ldots,n$. R_k is defined in definition 3.

Step 2 (Initial Clustering)

The similarity between the initial Equivalence Relation R_k and the others is computed. Similarity is presented by $S(R_k, R_l), l=1,2,\ldots,n$, the threshold of similarity of initial equivalence relation is assigned to β_2. General Equivalence Relation $R_k^{'}$ $(k=1,2,\ldots,n)$ is calculated, $R^{'}=\cap_{k=1}^{n} R_k$

Step 3 (Clustering again)

Initial Clustering results is merged if need. Firstly, calculates the similarity (or distances) between one cluster and others. Then, the two clusters with the largest similarity (or the smallest distance) are merged. The procedure is repeated until getting results we want.

For example, given that data stream set $D^n=\{d_1, d_2, d_3, d_4, d_5\}$， and the specified threshold of distance among data streams isβ_1, the initial equivalence relations are $R1=\{\{d1, d2, d3, d4\}, \{d5\}\}$， $R2=\{\{d1, d2, d3\}, \{d4, d5\}\}$， $R3=\{\{d1, d2, d3\}, \{d4, d5\}\}$ $R4=\{\{d1, d4, d5\}\ \{d2, d3,\}\}$， $R5=\{\{d1, d2, d3\}, \{d4, d5\}\}$. The similarity $S(R_k, R_l)$ between the initial Equivalence Relation R_k and the others is computed, $S(R1, R2)=3/4$， $S(R1, R3)=3/4$, $S(R1, R4)=2/5$, $S(R1, R5)=1/5$. If the threshold$\beta_2=2/3$, the similarity is calculated by same method. So the general equivalence relations are calculated, $D^n/R_1^{'}=\{\{d_1, d_2, d_3\}, \{d_4, d_5\}\}$, $D^n/R_2^{'}=\{\{d_1, d_2, d_3\}, \{d_4, d_5\}\}$, $D^n/R_3^{'}=\{\{d_1, d_2, d_3\}, \{d_4, d_5\}\}$, $D^n/R_4^{'}=\{\{d_1, d_2, d_3\}, \{d_4, d_5\}\}$, $D^n/R_5^{'}=$

$\{ \{d_1, d_2, d_3\}, \{d_4, d_5\} \}$ .Given $R' = \mathrm{I}_{k=1}^{n} R_k'$, so the clustering result is $D^n / R' = \{ \{d_1, d_2, d_3\}, \{d_4, d_5\} \}$.

Step 4 (Update Clustering results)

Dynamic k-means algorithm is called to update the clustering results.

Step 5

Output clustering results.

Step 6

{Clustering algorithm end}

The steps 1and 2of the algorithm need calculate the value ofβ_1,β_2, but the procedures of computing is not described for the limited space of the paper, please refer to the literature [5]. Step 4 is called the dynamic k-means clustering algorithm referred literature. The aalgorithm 1 brings up the initial clusters based on rough set. Because the data stream is changing gradually, the two data streams arriving adjacently are overlap mostly. Thus each time clustering, based on the previous clustering results, a small amount of iteration using k-means can get results.

4. Experiment results and analysis

Experiment environment: Intel Pentium 4 processor, 1GB of memory. Operating System: Windows XP. Algorithm is implemented by Microsoft Visual C + +. Data set used in experiment are the simulation data and real data sets.

4.1. Algorithm efficiency validation

Simulation data sets is used to compare Algorithm efficiency of RSBCAM with CluStream. The experiment results shown in Figure 2, which can be found that time cost by RSBCAM algorithm far less than CluStream algorithm in the condition of same data set. So the RSBCAM algorithm is more efficient than CluStream Algorithm.

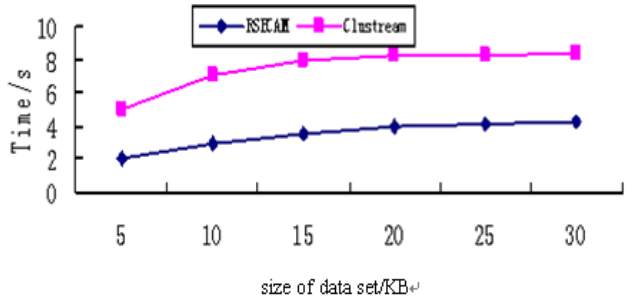

Fig.2 comparison of efficiency between RSBCAM and CluStream

4.2. Algorithm accuracy validation

Real data sets KDD-CUP-99 is used to compare the accuracy between RSBCAM and the CluStream. The results are shown in Figure 3. Figure 3(a) shows that the average accuracy of RSBCAM algorithm is better than 98% or more. Form Figure 3(b), we can see that RSBCAM algorithm have more precision

CluStream. Because multiple data streams RSBCAM algorithm have twice procedures, the initial clustering and the clustering again, and then dynamically adjust the clustering results.

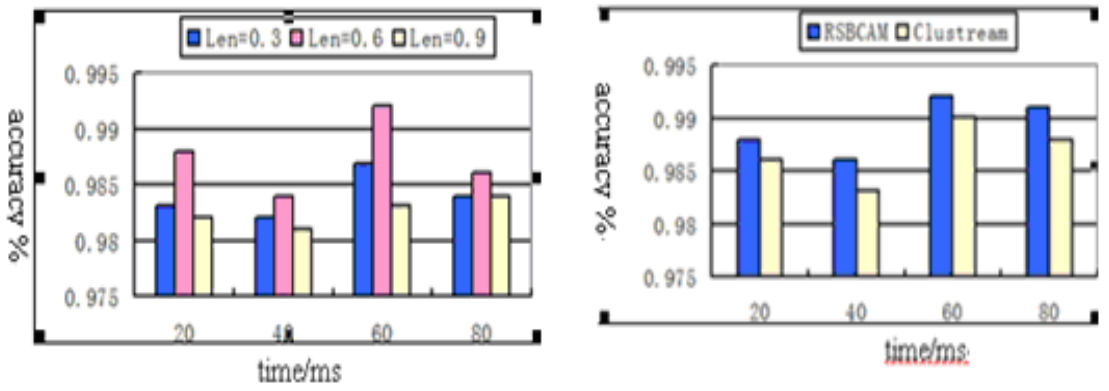

Fig.3. a) accuracy of RSBCAM algorithm; b}accuracy compared RSBCAM with Clustream.

5. Conclusion

The paper studies rough-set-based multi-stream clustering and proposes RSBCAM algorithm. The coupling degree is used to measure the similarity between the two data streams in term of change. It can reflect the structural characteristics of data streams. Experiments show that the algorithm can quickly and accurately identify clusters of data stream clusters, and have some practical value.

Acknowledgements

This work was financially supported by the Jiangsu Technology R&D Program Foundation (BE2009100), and supported by Huaiyin Institute of Technology Key Program Foundation (HGA0906)

References

[1]Golab L, TamerOzsu M. Issues in data stream management. *ACM SIGMOD Record,* 2003, P.5−14.

[2]Guha S, Mishra N, Motwani R, O'Callaghan L. Clustering data streams: Theory and practice. *IEEE Trans. on Knowledge and DataEngineering*, 2003,15(3):515−528.

[3] YANG Ning,TANG Chang-Jie, WANG Yue. Mining Evolutionary Events from Multi-Streams Based on Spectral Clustering. *Journal of Software*,2010, P.2395-2409

[4] Aggarwal CC, Han J, Wang J, Yu PS. A framework for clustering evolving data streams. In: Johann CF, Peter CL, Serge A,Michael JC, Patricia GS, Andreas H, eds. *Proc. of the 29th Int'l Conf. on Very Large Data Bases (VLDB 2003). New York: VLDB Endowment*, 2003,P.81−92

[5] ZHAO Ya-Qin, HE Xin WANG, Jian-Yu. An Effective High Attribute Dimensional Sparse Clustering. *Pattern Recognition and Artificial Intelligence*,2006, P.289-294