A Data Stream Subspace Clustering Algorithm

Xiang Yu[1], Xiandong Xu[1], and Liandong Lin[2]

[1]Department of Computer Science and Technology, Heilongjiang Institute of Technology,
Harbin, China
yuxpointfly@163.com
[2] Heilongjiang Province Key Lab of Senior- Education for Electronic Engineering,
Heilongjiang University, Harbin, China
1267013@qq.com

Abstract. The main aim of data stream subspace clustering is to find clusters in subspace in rational time accurately. The existing data stream subspace clustering algorithms are greatly influenced by parameters. Due to the flaws of traditional data stream subspace clustering algorithms, we propose SCRP, a new data stream subspace clustering algorithm. SCRP has the advantages of fast clustering and being insensitive to outliers. When data stream changes, the changes will be recorded by the data structure named Region-tree, and the corresponding statistics information will be updated. Further SCRP can regulate clustering results in time when data stream changes. According to the experiments on real datasets and synthetic datasets, SCRP is superior to the existing data stream subspace clustering algorithms on both clustering precision and clustering speed, and it has good scalability to the number of clusters and dimensions.

Keywords: data mining, data stream, subspace clustering, feature selection, dimension reduction.

1 Introduction

Recently, researches on data stream mining are motivated by more and more applications on continuous stream data, such as customer click streams, multimedia data, etc. By contrast with traditional data sets, data stream consists of a series of dynamic data objects which are massive and unordered[1,2,3]. Since not all data objects in data stream are maintained, it is necessary to be fault tolerant when clustering data stream. With the level of data collection technology increases and more and more characteristics data stream have, it is difficult to cluster data stream in high-dimensional space or cluster data stream in subspace effectively and efficiently. Clustering data stream in subspace needs to scan data sets several times, so it is almost impossible to process data stream with traditional subspace clustering methods.

HPStream[4] is a classic subspace clustering algorithm for high-dimensional data stream, Since it has several advantages such as fast clustering, high clustering precision, etc., HPStream is always used as the comparison to other algorithms on clustering speed and clustering precision. After data are partitioned into clusters, HPStream

H. Wang et al. (Eds.): ICYCSEE 2015, CCIS 503, pp. 334–343, 2015.

chooses clustering dimension with heuristic strategy. However, HPStream needs to determine the number of average dimensionality as parameter, which is hard to determine generally[5][6]. And there exists an obvious high-dimensional clustering problem. In 2007, Sun proposes GSCDS, a data stream subspace clustering algorithm based on grid which can cluster high-dimensional data stream[7]. In certain clustering subspace, GSCDS has high clustering speed and precision. When partitioning data space, the top-down partition method is adopted and then the subspace is related to clusters. GSCDS can identify clusters in different subspace with arbitrary shapes and the parameter of subspace need not to be predefined, and it has low computing complexity which is related to dimensionality of data space *d*, the number of grid cells *m* and the loop number of region partition, that is $O(dmn)$. However, when clusters in different subspace, GSCDS needs to run several times from top to down. In 2007, Park proposes a data stream subspace clustering algorithm based on grid[8][9], and it is based on, cell tree, a clustering algorithm on all dimensions. Park partitions data space into grid cells at different level and the grid cells are stored in a tree structure called sibling tree. In sibling tree, the clusters in each subspace can be easily found. There are three phases in this algorithm. In the first phase, a sibling tree is constructed by partition data space into grid cells at different levels. With data stream flows, the new coming data object falls into the corresponding grid cell according to its dimensional value, and the statistics of the grid cell is updated. In the second phase, dense grid cells are further partitioned. In the third phase, sparse grid cells are merged. The algorithm can cluster in all subspaces at different levels, but it needs to determine several parameters during the construction of sibling tree, and the parameters will further influence the final clustering results.

This paper proposes a new clustering algorithm SCRP(Subspace Clustering based on Region Partition), which can minimize the parameter influence on clustering and adopts down–top strategy. When partitioning regions, the distribution situations of data points on each dimension are considered. When clustering, the dense regions on each dimension overlap and then subspace clusters appear. Since the partition regions are formed on the basis of grid cells, SCRP has the advantages of grid clustering algorithms which can fast cluster and are not sensitive to outliers[10,11,12]. SCRP can find subspace clusters effectively and adjust the subspace cluster information according to the changes of data stream. When data stream changes, the region information will also change which will further influence the subspace clusters.

2 Definition

2.1 Concepts of SCRP

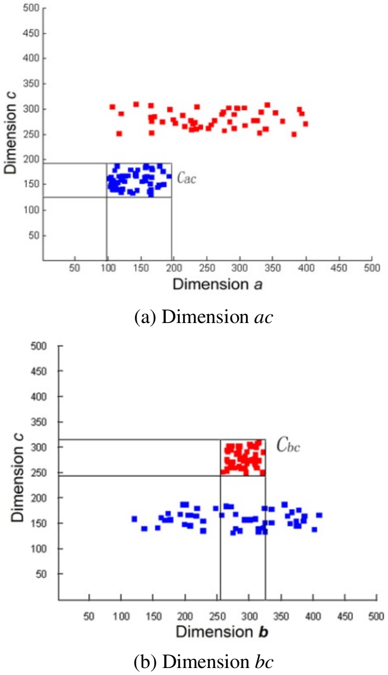
Dense grid Cell: The grid cells whose dense support exceed the dense threshold ρ.
Dense Region: The region consists of connected dense grid cells on each dimension.
Partition Regions: The region sets consist of dense regions and sparse regions which are formed by merging dense grid cells and sparse grid cells.

Region-tree : region–tree is a limited set formed of $k(k \le 2^d)$ subspace nodes. Region-tree has only one root node, and the number of sons involved in each node will not exceed *d*, where *d* is the dimensionality of data space.

2.2 Principle of Region Partition in SCRP

Suppose *D* is a set consists of *n* *d*-dimensional data points, $D = \{x_1, x_2, ..., x_n\}$, $D \subseteq \mathbf{R}^d$. Data point *x* is denoted as $x_i = \{x_{i1}, x_{i2}, ..., x_{ij}, ..., x_{id}\}$, where $j = 1, 2, ..., d$, and x_{ij} is the value of x_i on dimension *j*. When data stream flows, the slipping window is defined by time and the statistics information of each corresponding node in region-tree is updated according to the data in slipping window, such as the average value of data points in a region, etc. Then cluster in subspace according to requirements.

First, SCRP finds all dense regions on each dimension, and then merge connected dense regions and sparse regions to form partition regions. The set $R^1 = \{R_1^1, R_2^1, \ldots, R_d^1\}$ involves dense regions on each dimension, where R_i^1 denotes the set of dense regions on dimension *i*. For example, in 3-dimensional subspace, C_{ac} and C_{bc} are the clusters which are overlapped by dense region R_c^1 with R_a^1 in subspace *ac* and R_c^1 with R_b^1 in subspace *bc*, as shown in Figure 1.



(a) Dimension *ac*

(b) Dimension *bc*

Fig. 1. Two dimensional subspace clusters in three-dimensional space

Project the data in $D = \{x_1, x_2, ..., x_n\}$ on dimension *i*, the result is shown in figure 2. According to the distribution situation of data projection, when region partition on dimension *i* completes, the result of region partition is kept in R_i^1. Similarly, the result of ordered partition regions on other dimensions is kept in the corresponding sets. Finally, the dense regions on each dimension are kept in $R^1 = \{R_1^1, R_2^1, \ldots, R_d^1\}$.

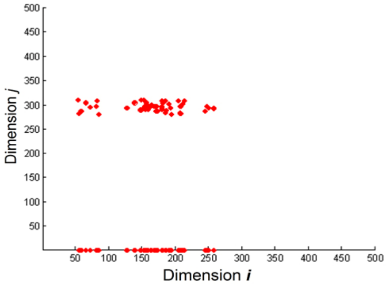


Fig. 2. Data projection on the *i*th dimension

2.3 SCRP Clustering

In SCRP, when clustering, the similarities of overlapped regions in subspace are compared with the specified threshold. If the similarity exceeds the threshold, the overlapped region is identified as a cluster of the subspace. The region similarity is denoted as $\mathrm{sim}(R_i^1, R_j^1) = | R_i^1 \cap R_j^1 |$, where $R_i^1, R_j^1 \in R^1$. In SCRP, a tree structure called region-tree is used to preserve subspace clusters. In region-tree, we can find clusters in all subspaces, and the sequence of dimensions is not important. Region-tree is constructed according to the predefined order $d_1 \rightarrow d_2 \rightarrow \ldots \rightarrow d_d$. Figure 3 shows a

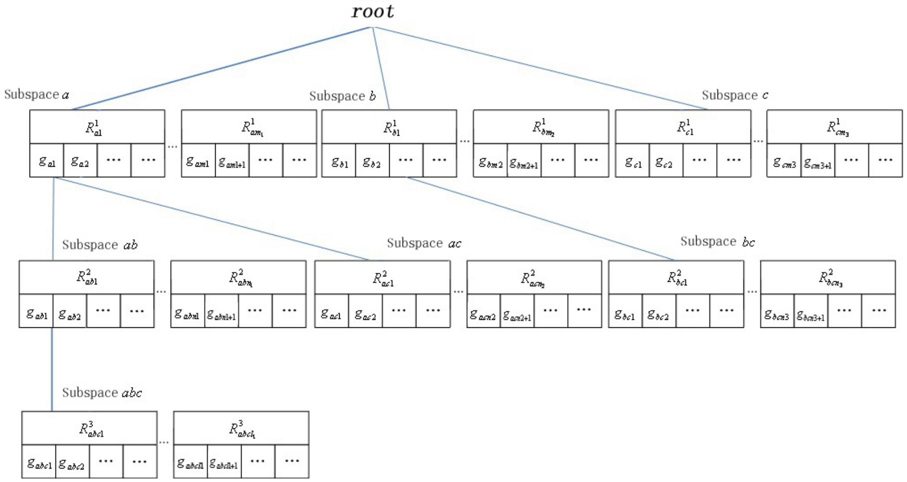


Fig. 3. Structure of Region-tree

region-tree in 3-dimensional data space, and *a*, *b*, *c* denotes three dimensions in data space respectively. In each data space, different regions and the grid cells involved in the regions are preserved in the region link structure. R_X^i denotes the region link in *i*-dimensional data space which records the basic information of each region in the data space, where subscript *X* denotes the region sequence number and the dimensions involved. The region in region-tree involves at least one grid cell. In grid cell g_Y , the statistics is recorded, and subscript *Y* denotes the sequence number of grid cells and the of the subspace. In region link, the statistics of each region can be computed by the statistics of the grid cells involved. The statistics of the region will continually be updated with the changes of the data points in the region.

3 SCRP Subspace Clustering

First, SCRP clustering mainly consists of two phases. In the first phase, the statistics information of each grid cell is recorded. Based on the statistics information, the partition regions are formed on each dimension. With the changes of data stream, the partition results and the corresponding statistics of regions is updated. When the density of grid cells changes, the regions formed by connected grid cells also change, and the partition regions need to be renewed.

The update process of region partition is as follows:

Sparse Region Split: When the density of sparse region exceeds the threshold, then split the sparse region into grid cells and reunite the grid cells into regions.

Dense Region Split: When the density of dense region falls under the threshold, split the dense region into grid cells and reunite the grid cells into regions.

In the second phase, construct region-tree, and record the information of each region and the grid cells involved. Based on the information, compute the similarity of overlapped regions and then cluster. The process of SCRP is as follows:

Algorithm: SCRP.

```
Input : $sp, \rho, \delta$ .
Output : $result$ .
$result$ = NULL;
$root$ = NULL;
$R^1 = \{R_1^1, R_2^1, \ldots, R_d^1\}$ ;
$PartitionSpace(\delta, d)$ ;
while ( $x_i$ arrives)
    update grid cell statistics information according to the value of $x_i$ on each dimension.
    for (each dimension $j$)
        region partition or update on dimension $j$
            preserve the partition results or update results in $R_j^1$
    endfor
```

```
endwhile
$root = Construct\_RT(R^1)$;
$result = SP\_Clustering(root, sp)$;
return $result$;
```

In the algorithm, *sp* denotes the clustering subspace, ρ, δ denotes the dense threshold and the partition parameter of grid cell respectively, *result* preserves the clustering result, R^1 preserves the partition regions on each dimension, and *root* denotes the root of region-tree.

SCRP first partitions *d*-dimensional data space with parameter δ, and updates the statistics information of grid cells according to the arrival of data points. When clustering in subspace, the connected grid cells on each dimension whose densities exceed threshold ρ are merged, further the partition regions are formed and preserved in R^1, then construct region-tree and cluster in the subspace *sp*.

4 Experiments Analysis

In order to analyze the clustering speed, clustering precision and scales, KDD-CUP99, a real-world data set of MIT, together with a synthetic data set containing 500000 60-dimensional data objects generated by a data generator are used. In the synthetic data set, the domain of data point on each dimension is [0,500], and each data point is involved in one cluster or exists as an outlier. In data space, the outliers in the synthetic data set are less than 5 percent of the total data points and are uniformly distributed.

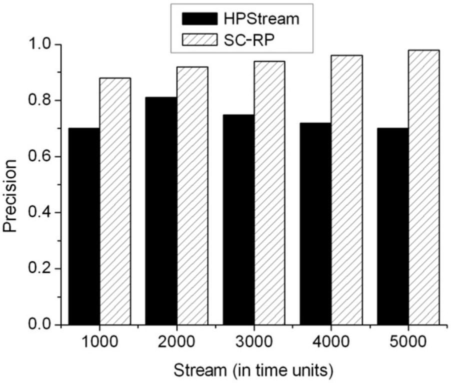
4.1 Quality Evaluation

Commonly, when clustering data stream at different time or in different subspace, the results will be different, and it is hard to evaluate the clustering quality when simulating data stream with the data in synthetic data set. Since CLIQUE is a clustering algorithm with high precision, its clustering result is always used as the measure to evaluate the clustering quality of other algorithms.

In the experiments, the clustering result of CLIQUE is used as the precision measure when compare the clustering quality between SCRP and HPStream. CLIQUE runs several times with different partition granularities, and the best clustering result is used as the precision measure. When simulate data stream changing, CLIQUE needs to run several times to guarantee the clustering precision. In order to test the validity of the algorithm, the synthetic data set is divided into two parts to simulate the wave change of data stream.

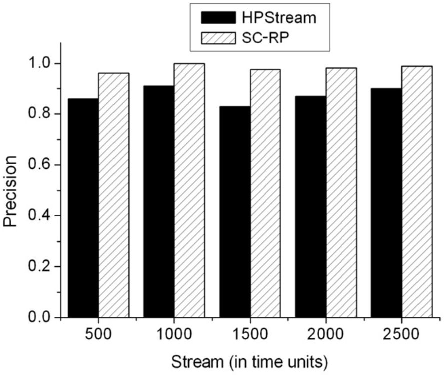
In this experiment, parameter δ is set $\delta = 20$ in SCRP, when the new coming data points increase to 1000, the region density and the region-tree need to be updated because of the density decay. The parameter of HPStream is set $l = 20$, where l denotes the average dimensionality. The comparison results of clustering precision are shown in figure 4 and figure 5. As can be seen, the clustering quality of SCRP is superior to HPStream.

When data stream changes, the historical information decays to reduce the influence on SCRP clustering. Other than HPStream, SCRP preserves historical information but reduces its influences gradually. According to the changes of data stream, SCRP updates the statistics information of regions and gets accurate clustering results. When wave changes coming, the stability of SCRP is superior to HPStream. Compared with HPStream, SCRP does not need to divide data into clusters and provide the average dimensionality in advance, in addition, SCRP does not have too much requirements on parameters.



KDD-CUP99, stream speed 200/s

Fig. 4. Comparison of precision

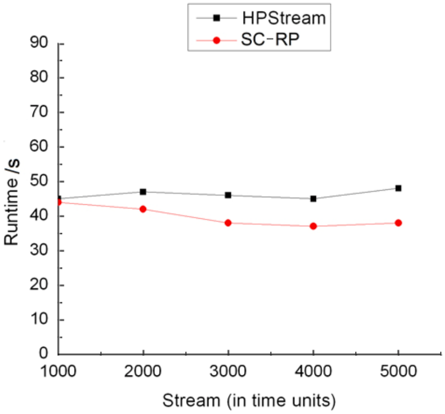


Synthetic data set, stream speed 200/s

Fig. 5. Comparison of precision

4.2 Runtime Evaluation

In the experiment, the parameter of SCRP is $\delta = 20$, when the new coming data points increase to 1000, the region density and the region-tree need to be updated because of the density decay. The average dimensionality is set $l = 20$. As shown in figure 6, compared with HPStream, SCRP needs less time, and it becomes more obvious when the region-tree is formed and tends to stable.

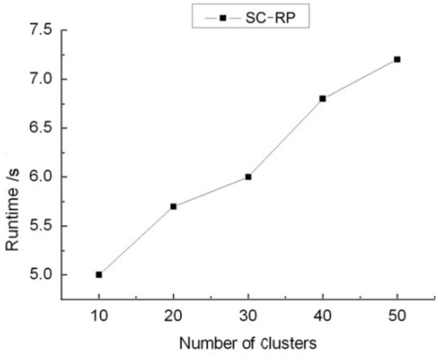


Synthetic data set, stream speed 200/s

Fig. 6. Comparison of runtime

4.3 Scalability Evaluation

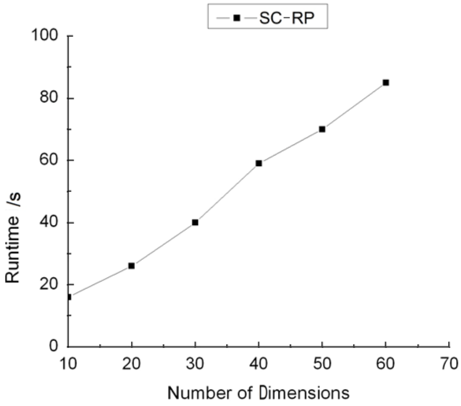
In the experiment, the parameter of SCRP is $\delta = 20$, when the new coming data points increase to 1000, the region density and the region-tree need to be updated because of the density decay. The experiment aims to test the scalability of SCRP when the data dimension and the number of clusters change. As shown in figure 7, SCRP has good scalability to the number of clusters, and its runtime is almost linear to the number of clusters in subspace, as the number of clusters increases, the runtime also increases.



Synthetic data set, stream speed 100/s

Fig. 7. Counts of clusters in subspace

SCRP is different to CLIQUE, its performance can hardly be influenced by the dimensionality, and it has good scalability to the changes of dimensionality. As shown in figure 8, the runtime of SCRP is almost linear to the dimensionality, as the dimensionality increases, the runtime also increases.



Synthetic data set, stream speed 100/s

Fig. 8. Change of dimensions

5 Conclusion

In this paper, a region-tree structure is designed to preserve the changes of data stream, and SCRP, a data stream subspace clustering algorithm based on region partition is proposed. Compared to the current data stream subspace clustering algorithms, SCRP can cluster in subspaces at all levels and costs less time. In addition, SCRP can hardly be influenced by outliers and it can record and adjust the clustering results according to the changes of data stream in region-tree. The experiments on real data set and synthetic data set show that SCRP has good effectiveness and applicability, and it is superior to traditional data stream subspace clustering algorithms. In the future, we will extend our work on how to reduce the influences on clustering quality when the grid partition granularity changes.

Acknowledgments. This work is supported by the science and technology project 12531542 of the education office of Heilongjiang province. We thank professor Qi haoliang for enlightening discussions.

References

1. Ling, C., Lingjun, Z., Li, T.: A clustering algorithm for multiple data streams based on spectral component similarity. Information Sciences 183(1), 35–47 (2012)
2. Weiguo, L., Jia, O.: Clustering algorithm for high dimensional data stream over sliding windows. In: Proc of the 10th Int. Conf. on Trust, Security and Privacy in Computing and Communications, pp. 1537–1542. IEEE, Piscataway (2011)
3. Halkidi, M., Koutsopoulos, I.: Online Clustering of distributed streaming data using belief propagation techniques. In: Proc of the 12th Int. Conf. on Mobile Data Management, pp. 216–225. IEEE, Piscataway (2011)

4. Aggarwal, C., Han, J., Wang, J., et al.: A framework for clustering evolving data streams. In: Proc of the 29th Int. Conf. on VLDB, pp. 81–92. Morgan Kaufmann (2003)
5. Parsons, L., Haque, E., Huan, L.: Subspace clustering for high dimensional data: A review. ACM SIGKDD Explorations Newsletter 6(1), 90–105 (2004)
6. Yihong, L., Yan, H.: Mining data streams using clustering. In: Proc of the 4th Int. Conf. on Machine Learning and Cybernetics, pp. 2079–2083. IEEE, Piscataway (2083)
7. Yufen, S.: Research on clustering algorithm based on grid. Huazhong University of Science and Technology, Wuhan (2006)
8. Park, N.H., Lee, W.S.: Cell tree: An adaptive synopsis structure for clustering multi-dimensional on-line data stream. Data & Knowledge Engineering 4(3), 1–22 (2007)
9. Park, N.H., Lee, W.S.: Grid-based subspace clustering over data streams. In: Proc of the ACM Conf. on Information and Knowledge Management, pp. 801–810. ACM, New York (2007)
10. Yanwei, Y., Qin, W., Jun, K., et al.: An on-line density-based clustering algorithm for spatial data stream. Acta Automatica Sinica 38(6), 1051–1059 (2012)
11. Dutta, B.R., Angelov, P.: Evolving local means method for clustering of streaming data. In: Proc of the 2012 Int. Conf. on World Congress on Computational Intelligence, pp. 1–8. IEEE, Piscataway (2012)
12. Lingjuan, L., Xiong, L.: An improved online stream data clustering algorithm. In: Proc of the 2nd Int. Conf. on Business Computing and Global Informatization, pp. 526–529. IEEE, Piscataway (2012)