Research Journal of Applied Sciences, Engineering and Technology 6(2): 258-264, 2013
DOI:10.19026/rjaset.6.4067
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: July 12, 2012 Accepted: September 08, 2012 Published: June 10, 2013

Research Article

A Data Stream Clustering Algorithm Based Extension of Grid and Density

Yang Yongbin and Ding Mingyong
College of Computer Science and Information Engineer, Chongqing Technology and Business University, China

Abstract: This study focuses on the summary data structure design and optimize the method of calculation of the mesh density and how to effectively deal with the problem of boundary points, combined with the sliding window mechanism and suggest improvements based on the mesh density of the data stream real-time clustering algorithm framework and the various parts of concrete realization of the algorithm.

Keywords: Clustering algorithm, data stream, extension, grid

INTRODUCTION

Grid density calculation: Density value in the part of the existing data stream clustering algorithm is defined as the number of objects in the cell, although such a convenient handle and correctly and efficiently, but accuracy is not enough, because it does not take into account the influence of the boundary points. If a point has 2 characteristics, this method will lead to this point is divided into one of this does not include its category, the clustering results are not accurate. Figure 1 shows a spherical primitive class, if the number of data points in the grid as the density values are calculated, the results shown in Fig. 2. Obviously, the clustering results are not accurate, a part of at the edge of the point (B, C units of the point), is likely to be discarded as "noise" data, because this part of the algorithm in the combined grid will be less than density discard threshold grid as sparse grid, leaving only the grid cell to meet the threshold conditions in unit A will be retained as part of the clustering results, B, C, the point would be to do as a noise delete.

Denclue use of a data on its impact around the idea put forward the concept of influence function. Single Shimin (2006), who studied the relationship between data objects and their spatial data around on the basis of a known as the "contribution" is defined, the data objects in a grid by calculating the contribution to this grid degree of grid density, but the time complexity of large and fast dynamic data flow changes is contradictory. Yong (Han and Kamber, 2006; Arlkerst *et al*., 1999; Hinneburg and Keim, 1998; Wang *et al*., 1997; Agrawal *et al*., 1998; Sheikholeslami *et al*., 1998), who proposed a dynamic grid data stream clustering algorithm, proposed the concept of the extended grid cell, the 4 sides of the original grid cell

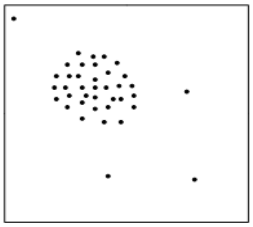

Fig. 1: Original class of globe

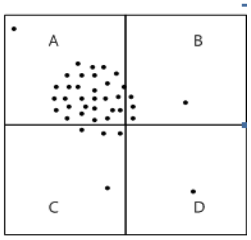

Fig. 2: Original grid clustering schematic

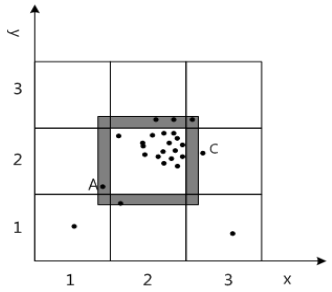

Fig. 3: Extended grid unit

length extended to 4 weeks for the 1/4 of the grid side length shown in Fig. 3, the grid cell is extended. In the

Corresponding Author: Yang Yongbin, College of Computer Science and Information Engineer, Chongqing, Technology and Business University, China

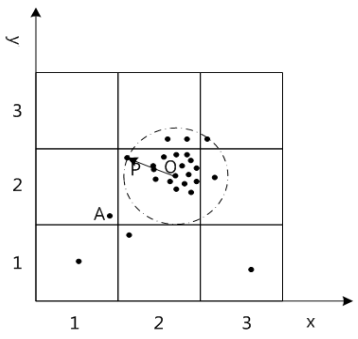

Fig. 4: Grid extension diagram in

calculation of the density of the grid cell, coupled with the point of the grid cell around 1/4 side of the gray area, the degree of impact on the grid.

Although this method has been proved through experiments, taken to extend the 1/4 side length of the extension methods, the clustering result is valid, but with the changes in the data set, this monotonous expansion is clearly inappropriate. Dynamic variability of the data stream, the algorithm and also with the changes in the data stream and dynamic adaptability. If the data points

Distribution shown in Fig. 3, which shows, the point A on the grid unit is obviously should not be taken into account, but in the extended grid of side 1/4 long, the point A into consideration and if so more than 1 point in the calculation of grid cell density, add these points have influence, in the case of such a distribution, a grid did not contribute to the point, but consider the density of the grid calculation, which is clearly undesirable and point C in Figure clearly should be considered within the extended range of the grid, but discarded the point C, showing that this extension methods there are some deficiencies.

This topic based on the concept of functions and improves the idea of expansion of the grid, a new method based on the extension of the grid edge to calculate each grid cell density.

This improved extension, two-dimensional space, based on the grid cell center dot grid cell away from the center of the distance between the farthest point and the center (g-uc) radius of a circle, such as in Fig. 4 shows the coordinates of all points in the grid cell to take the center point, to get the center is the center of the circle is point O, as the center point O to point O to point P from O farthest grid the distance between the radius circle.

The benefits of doing so is to avoid the problems described above in the expansion of the grid cell boundary method, some points on the grid cell has been incorporated into the consideration affect the calculation accuracy of the mesh density, which laid an effective foundation for further clustering.

DENSITY THRESHOLD

The researchers found that for the data in the database, clustering the results they get are rules to follow, the larger the similarity based on the similarity function or data with characteristics similar to the data object, due to the distance between the smaller will be closer to those close to the object point set from the area denoted by dense region; the contrary, the more dispersed the data, there is no common characteristics of the data, far away from each other object called sparse regions. Classical density-based clustering method, the density threshold (Agrawal *et al*., 1998) the idea to achieve the purpose of clustering. Learn from this thinking, in the data stream clustering algorithm based on the density of the grid, due to the need to determine which is a dense grid, which grid is sparse, so, you need to set the decision parameters to distinguish between different grids in order to enter the next merger.

In many existing algorithms require the user to set require the user to relevant background knowledge and on the basis of the relevant attributes of the cluster system to a better understanding of in order to make better choices, which limits the poly class of algorithms suitable for the specific range of users, it reduces the practicality of the algorithm. Because most users do not have the relevant background knowledge and field skills and set parameters are appropriate to determine whether the results of the last algorithm runs in line with the facts. In other words, if the parameters are set incorrectly, the results are likely to cluster is another deviation or error, such a result and then users with the actual, is clearly undesirable. Fast dynamic changes in the characteristics and the data stream, apparently a fixed set of a certain threshold or user man-made changes in parameters, it is difficult to adapt to the rapid flow of data clustering.

Existing classical density clustering algorithm ideologically dynamic change of the adaptive data to calculate the threshold g-Minpts, the threshold calculation method is effective to avoid artificial obstacles set up and as dataflow into the dynamic change in order to achieve real-time clustering. Formula as described below.

Density of mean formula:

$$g\_average = \frac{\sum_{i=1}^{k} g\_density_i}{k}$$

(The number of which K non-empty grid)

The density threshold formula:

$$g\_Minpts = \frac{g\_average + g\_max}{2}$$

$$g\_Low = \frac{g\_average + g\_min}{2}$$

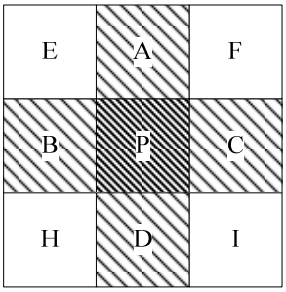

Fig. 5: Adjacent grid schematic diagram in

Density g-max, the maximum value, minimum value is g-min, i.e., scanning the B data of the inflow window, the max and min values of data in the log window, window update, g-max, g-average also will be updated, so that g-Minpts can adapt to the dynamic changes of the data stream.
Dense grid is defined as:

g-density> = g-Minpts (Grid cell density threshold)

Sparse grid is defined:

g-density< = g-Low

Medium grid is defined as:

g-Low<g-density<g-Minpts

Density grid merge: The definition of an adjacent cell if and only if the 2 units overlap the edge or coincidence point. The exposition of the mathematical language is, if the absolute value of the 2 cell dimension index number subtract the result is equal to 1 or equal, then these two cells are adjacent (Sheikholeslami *et al*., 1998).

The d-dimensional data space, any cell and 2D cell adjacent. Figure 5 for the 2-dimensional grid structure and grid P and grid A, B and C and D, is adjacent grid E, F, H, I and grid A, B and C and D, respectively.

Among them, the more dense slash describe the grid such as P for the intensive, more sparse slash grid such as A medium grid cell, blank unit E represents the sparse grid.

Grid cluster: If P is a dense mesh, containing P and P adjacent to the dense grid with the medium grid and together is called a cluster. Grid cluster core is defined as the intensity of the largest grid cell cluster. Figure 5 shows, the P as the core of a cluster, denoted by (P, A, B, C and D).

Shown in Fig. 6, the cluster center as the P1 cluster and the cluster center for the P2 cluster, 2 clusters have not yet merged, among them is connected by the grid A, as shown here, grid A is dense grid, namely g-density (A)> = g-Minpts, then the grid clusters P1 and grid

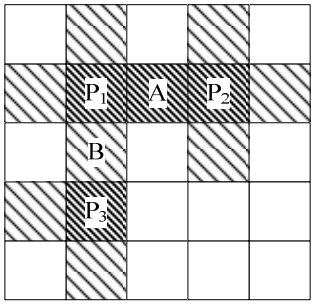

Fig. 6: Grid cluster connectivity diagram

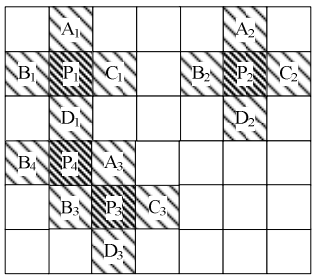

Fig. 7: Cluster merger schematic

cluster P2 cluster connectivity and high; the contrary, the cluster center for the P1, P3 on the cluster and the cluster center between clusters connected by the grid B medium density grid, that is g-Low <g-density (B) <g-Minpts, called cluster P1 and cluster P2 cluster connectivity.

Merge method of the grid is divided into two types, 1 is the merging of adjacent grid; a cluster are merged based on the cluster connectivity. This draws on the experiences of the traditional density-based clustering methods, the classical density algorithm is used, directly or indirectly density can reach is the distance between the data objects, the similarity concept clustering, where the merge the similarity of the grid, the grid density in regions with low in regions with high mesh density separated, the final result can make the cluster within the data object is quite similar to the cluster and cluster data object similarity between lower, which is the clustering algorithm targets.

Adjacent cells merge: Single grid cluster is by intensive grid unit began looking for a cell adjacent to this cell and merge from. As mentioned earlier, in a d-dimensional data space, single grid will exist around the 2d cell and its neighbors and if this unit together with its adjacent cell to meet the terms of the merger of adjacent cells, that is grouped together into a whole, this overall cluster as the grid.

For convenience, the following examples of 2-dimensional space, as shown in Fig. 7, grids P1 dense grid, that is, its density to meet P1-density> = g-Minpts, from P1 to begin scanning the surrounding grid, as long as non-sparse unit put this grid integrated into the

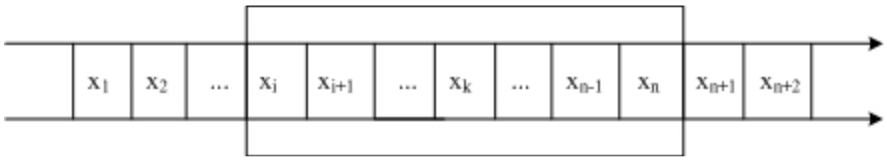

Fig. 8: Sliding window

cluster of P1 where P1 around four grids are for the middle unit, i.e., their density to meet g-Low< {A1-density B1-density, C1-density, D1-density} <g-Minpts, so I finally get with P1 as the core unit cluster (P1, A1, B1, C1, D1). And so on, you can get P2 for the core unit of the cluster (P2, A2, B2 and C2, D2) and P3 as the core unit cluster (P3, A3, B3, C3, D3) is.

Grid cell cluster merger: Unit cluster refers to merge with the cell clusters do not coincide with clusters between adjacent grids allow these 2 clusters connected to each other and then determine whether defined above cluster connectivity of the level of these clusters are normalized and replaced by the cluster core unit.

This study presents a grid between the cluster merger conditions, connectivity to the density threshold (g-Minpts) above can be grouped, that is, the only cluster connectivity and high in order to merge clusters.

On the basis of Fig. 8, assuming that the newly arrived data point is mapped to the grid structure and intensive unit P4 or P4-density> = g-Minpts and its adjacent cells B4, shown in Fig. 9 P4 of this unit, between the cell clusters and cell clusters can be connected to each other, when necessary according to the level of connectivity to determine these 2 clusters whether incorporated into a cluster.

Unit cluster P3 and unit cluster P4 is relying on the 2 at the same time the adjacent grid cell A3 and B3 connect, according to the terms of the merger, both A3 and B3 must be one becomes a dense unit to meet the merge premise, both moderate units, namely g-Low< {A3-density, B3-density} <g-Minpts, that is the connectivity between clusters is low, cannot merge. Merged at the same time, grid cell A3 and B3 at the same time these 2 clusters contain, re-analysis of these 2 units in the end belong to which cluster here P3-density>P4-density A3 and B3 attributable to the core P3 clusters, which produce the cluster (P3, A3, B3, C3 and D3) and cluster (P4, D1, B4) (here for the time being not consider the impact of the core unit of P1 clusters); On the contrary, the resulting cluster (P3, C3, D3) and cluster (P4, D1, B4, A3, B3,).

By the merger of the above shows, the connection between P1 and P4 unit is D1, if the new data arrived, D1, by the middle grid becomes dense mesh namely meet D1-density g-Minpts, P1 and P4 where the cluster connectivity is high, to satisfy the merger conditions. On this basis, the need to determine, after the merger of the cluster core unit, is to find the mesh density in the

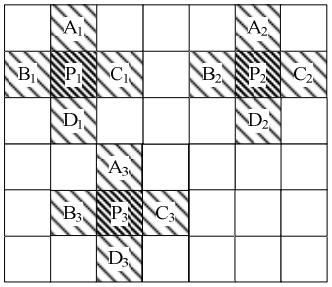

Fig. 9: Grid cluster formation

unit D1, P1 and P4 is larger as a new core unit, assuming that the density of the three relations for D1-density {P1-density, P4-density} and P4-density>P3-density new cluster at the core of the D1 unit, denoted as clusters (D1, A1, B1, C1, B4, the A3, B3,).

And so on, are now discussing a merger between the 3 clusters, D1, A3, B3, 3 grid cells are moderate unit g-Low< {D1-density, A3-density, B3-density} <g-Minpts, the core unit, followed by P1 between P3 and P4 cluster 2 cannot be combined, if the meet P1-density <P3-density <P4-density, then the resulting 3 clusters followed by cluster (P1, A1, B1, C1), cluster (P3, C3, D3) and cluster (P4, D1, B4, A3, B3); On the contrary, D1, A3, B3, 3 mesh-intensive unit {D1-density, A3-density, B3-density} > = g-Minpts, that is to say three clusters of between both satisfy the merger conditions can be combined into a new cluster, you need to determine a new core unit, i.e., to find out the density of the maximum value of cell D1, A3, B3, P1, P3, P4, suppose D1-density {A3-density, B3-density, P1-density, P3-density, P4-density}, is denoted by D1 as the core unit cluster (D1, A1, B1, P1, C1, B4 and P4, the A3, B3, P3 and the C3 and D3).

SLIDING WINDOW MECHANISM

The data stream has a high-speed flow, rapid change and potentially unlimited features, data stream processing algorithms usually need to select a time period depending on the application data processing. The data flow model is divided into 3 models based on different time length of the form: 3 of the landmark model (model), the sliding model (the sliding model) and a snapshot of the model (model) (Aggarwal *et al*., 2004). Landmark model is defined as the data collection period of time, that is, from a point data set to the current moment came and then this piece of data sets for processing; sliding window model is to deal with the latest data to reach the sliding window. The

snapshot model is processing between the 2 pre-defined time nodes within the data segment. Landmark and sliding window, the 2 models can continue to deal with newly arrived data easier to use in a real study and scholars of all ages.

The sliding window can handle a moment from the past to the current time within the scope of data to avoid outdated data and data flow analysis and statistical approximation of the limited storage space tool. The sliding window is divided into a sliding time window (time-based the sliding window), the window size is defined by a time interval and count the sliding window (count-based the sliding window), that is, how much the definition of the amount of data contained by the window. Therefore, the sliding window technique to reduce the amount of data the algorithm needs to be addressed, mining changes in data streams, so as to achieve real-time accurate on-line clustering results indispensable technical support reserve (Zhu, 2006; Cao *et al*., 2006).

As shown in Fig. 9, the algorithm of this study is based on the count-window model; you can use a FIFO queue that the window size remains unchanged and always hold the N objects. Time increases, new data points continue to slide into the left side of the window, the old data objects continued to slide out from the right end of the window. Whenever new data into the window, you can make a rapid update of the grid unit Eigen value or density threshold; old data off the window, but also the instantaneous impact Excluding these historical data, in order to achieve growth the amount of type of clustering results to perform the update.

Each entry in the window one data point, an update on the cluster, although the updates are very fast, but consumption will be relatively large. Therefore, in implementing the relevant algorithms, use the window the sliding Win-STEP step to perform the update operation on the cluster update granularity from a change for the Win-STEP, sliding Win-STEP step, then the Win-STEP new (old) data points slid into (out clustering execution time) window after the update operation.

EGDDSC algorithm: The benefits of grid density-based clustering algorithm is the ability to fast data clustering and the shape of the limitation of the clustering results can be non-spherical shape of the cluster. This subject in the existing data stream clustering algorithm is proposed to improve the idea of an extension of the grid, by the Statute of the transformation data corresponding to the corresponding grid and then use the adaptive density threshold on the data objects in the grid clustering.

Of this study is based on an extension of the grid density, denoted EGDDSC algorithm, refer to the related idea of the classic B algorithm, algorithm B is

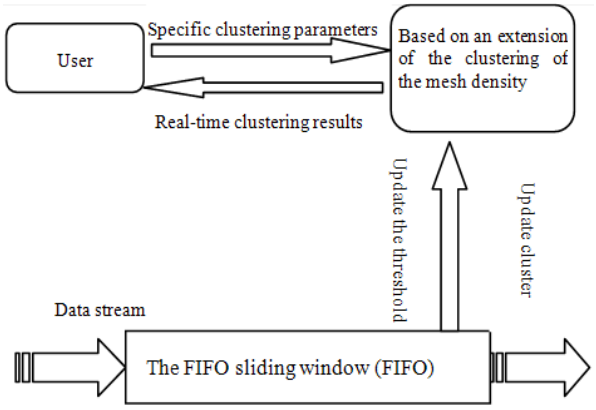

Fig. 10: Data stream clustering framework diagram

the online stage is mainly responsible for keeping the summary of the data stream by the online and offline 2 phases, with the corresponding overview of the structure stored, processed and real-time clustering data update window. Offline stage is based on the user's interest, according to the needs of users from the online stage to save the results to extract the appropriate information. Offline stage are independent and can use any of the processing model or algorithm to meet user needs, so the article focuses on the process of online clustering algorithm to get real-time clustering results.

Of this study is based on an extension of the grid density, denoted EGDDSC algorithm, refer to the related ideas of the classical algorithm, the algorithm is the online stage is mainly responsible for keeping the data stream summary information from online and offline 2 phases and the corresponding summary structure storage, processing and real-time data update window clustering results. Offline stage is based on the user's interest, according to the needs of users from the online stage to save the results to extract the appropriate information. Offline stage are independent and can use any of the processing model or algorithm to meet user needs, so the study focuses on the process of online clustering algorithm to get real-time clustering results.

The algorithm for the overall framework: Shown in Fig. 10, the use of window technology, 1st in, 1st out of the window, data window, the data transformation Statute, mapped to the grid space and removes the "noise", finishing the incomplete data for the start of the implementation clustering algorithm and updates the clustering algorithm as a foundation. Clustering algorithms, the calculation method used in this study based on the extension of the grid density threshold, as the data changes and update the threshold has been reached to make adaptive to the rapid changes in the data stream, while using an improved grid merge method, using the low area of the grid density in regions.

The results obtained can make a higher similarity of the cluster within the similarity between the cluster and the cluster is relatively low, in order to effectively achieve real-time clustering.

The online real-time clustering algorithm principle is the number of object count sliding window, with the arrival of the data, statistics, sliding window, when the initial number of clusters is reached, start the initialization of the clustering algorithm. Data continuously into the FIFO by the sliding window mechanism, the head gradually into the new data, slide the tail of the old data out of the window, sliding window size, here is determined by the number of data objects to the window the data window size, whenever the number of data sliding window equal to the preset Win-STEP value update processing to update the threshold value and updates the clustering results. Set the data space for the n-dimensional, the algorithm started, you need to set the number of each dimension of the grid Grid-K n-dimensional (Grid-K) n grid cell. The algorithm is described as follows:

Input: Data stream X {x1, x2, ..., xi, ...}, Grid-K, Win-N, Win-STEP.

Output: Real-time clustering results

1 Specified parameters, divided into n-dimensional grid space structure
2 The sliding window capacity to count win-n = 0, sliding objects the number win-step = 0
3 The while new data xi = (xi1, xi2, ..., xin) into the sliding window do
4 mapped to the newly added data object xi to the corresponding grid cell g, xi->xi ';
5 win-n + +
6 if win-n = Win-N, the calculation of the initial threshold and run the clustering initialization algorithm
7 if win-n>Win-N
8 win-step + +
9 while win-step = Win-STEP
10 to calculate and update the threshold to perform the update clustering algorithm
11 step = 0
12 end while
13 end while

Clustering initialization algorithm: As mentioned above, the overall algorithm is described in Step 6 is the clustering initialization, initialize the clustering algorithm The main idea of steps, within the window data Statute transform the defined grid space. Calculated the corresponding grid Eigen value and then calculated according to the grid Eigen value initial threshold g-Minpts, g-Low. Based on the threshold, record grid properties, dense, medium, or sparse cell. Then, as a dense cell cluster core unit analysis of the adjacent grid and grid merge merger in accordance with the above described method; cluster has been formed on this basis, in accordance cluster consolidation method, under the conditions that satisfy the cluster merge cluster implementation of the merger. Based on the above operation, the final clustering results of the data within the window.

Update the clustering algorithm: The basic idea is to update the clustering algorithm, the use of the count of the sliding window mechanism, by the above, the window size is measured by the number of the transaction object within the window, set window size Win-N, window can accommodate the number of data constant Win-N. In the implementation of the algorithm, in order to ensure that the speed of real-time clustering and to ensure that the simplicity of the algorithm here to update when the window every the mobile Win-STEP times the amount of data and fast processing. That is updated, the window will enter Win-STEP new data window also retained Win-N-Win-STEP, old data, then the need to re-calculate the density threshold, the threshold update.

New data, new intensive unit and the intensive unit does not meet the terms of the merger of any grid clusters, put this intensive unit as new cluster core units independently; sparse unit, no further treatment. Upgrade for the middle cell, until it came in the new data, then merge operation. Through the above methods of operation to achieve a quick update on the clustering and get real-time clustering results.

Algorithm complexity analysis: Assumptions to be processing the data set the number of data points is N, window size Win-N, the data space is d-dimensional algorithm is a mesh, set each dimension k grid, here the grid size is fixed, so the time complexity is constant, then the data corresponding to a specific grid, the time complexity of O (d*Win-N), calculate the density threshold, the 1st scan is performed on the grid within the window Eigen value complexity O (Win-N) grid in the merger, a number of adjacent cells of the grid cell up to 2D, so all data corresponding to the unit in the window to merge the highest number of Win-N*2d, we can see, clustering the time complexity of the algorithm is approximately O (d*Win-N) means that the algorithm and the space dimension and the window number of data objects into a linear relationship.

Each dimension is divided into k grid, so up to save kd cell in memory, which means that the space complexity varies with the change of the space dimension and mesh size. When the number of data objects in the cell degradation to zero, which means that the grid cell is empty, the algorithm to put the memory of this unit release to avoid a waste of space occupied the effective recovery of memory.

Table 1: Experiment environment

CPU	Pentium (R) 4.00 GHz
Memory	1G
Operating system	Windows XP SP2
Development tools and language	Microsoft visual C++6.0 C language

Update algorithm, the Win-STEP a data object in the window on the clustering of sexual update operation, rarely need to operate in the updated data object, processing time is shorter, saving overhead, thus achieving the online the goal of effective real-time clustering.

ALGORITHM ANALYSIS AND EXPERIMENT

This chapter from the experimental point of views the analysis and verification algorithm performance and actual effectiveness. The validation of clustering accuracy and efficiency of the algorithm simulation and real data. The simulation data verified the number of steps of the algorithm in a sliding of the sliding window to set the degree of clustering results of the algorithm; clustering algorithm is indeed feasible and effective through the real data sets to verify, so the simulation data on the basis of real data sets.

Experimental environment and data:
Experimental operating environment: Algorithm operating environment as shown in the Table 1.

The data set: The data used in this experiment, including simulation data sets and real data sets.

The simulation data sets: The multi-threaded ideological programming, random constantly changing data to simulate the dynamic fast-moving stream data; set the speed of data flow to.

The real data sets: Algorithm design, to be verified on real data sets, the real data sets used in this study is Forest, Cover Type this dataset in the UCI website, recording a total of more than 50,000, each record contains 10 properties.

Experimental analysis:
The Win-STEP sliding window experiment: The proposed algorithm is based on the count sliding window mechanism, update number of the sliding of the window Win-STEP set over the number of data objects in the slide on the clustering will cluster results have a greater impact. The experiment is to test by set different Win-STEP, Win-STEP clustering accuracy and clustering speed.

Evaluation of clustering quality is average purity (Average purity) method, the purity of the calculation of form shown in the following formula:

$$pur = \frac{\sum_{i=1}^{N_c} \frac{|C_i^d|}{|C_i|}}{N_c} \times 100\,\%$$

where,

N_c = The cluster-cluster number

$|C^d_i|$ = The number of i clusters with the number of objects of the cluster class label

$|C_i|$ = Said that the number of objects of cluster number i

REFERENCES

Aggarwal, C., J. Han, J. Wang and P. Yu, 2004. A framework for projected clustering of high dimensional data streams. Proceeding of the 30th International Conference on Very Large Data Bases. Toronto, Canada, pp: 852-863.

Agrawal, R., J. Gehrke, D. GonoPulos and P. Raghavan, 1998. Automatic subspace clustering of high dimensional data for data mining applications. Proceeding of the ACM SIGMOD International Conference on Management of Data. Seattle, WA, US, pp: 94-105.

Arlkerst, M., M.M. Breunig, H.P. Kriegel and J. Sander, 1999. OPTICS: Ordering points to identify the clustering structure. Proceeding of ACM SIGMOD International Conference on Management of Data. Philadelphia PA, pp: 49-60.

Cao, F., M. Ester, W. Qian and A. Zhou, 2006. Density-based clustering over an evolving data stream with noise. Proceeding of the 6th SIAM International Conference on Data Mining. Bethesda, MD, pp: 328-339.

Han, J. and M. Kamber, 2006. Data Mining: Concepts and Techniques. 2nd Edn., Morgan Kaufmann Publishers, pp: 196-220, ISBN: 1-55860-901-6.

Hinneburg, A. and D. Keim, 1998. An efficient approach to clustering in large multimedia databases with noise. Proceeding of the 4th International Conference on Knowledge Discovery and Data Mining. NY, USA, pp: 58-65.

Sheikholeslami, G., S. Chatteqee and A. Zhang. 1998. Wave cluster: A multi-resolution clustering approach for very large spatial databases. Proceeding of the International Conference on Very Large Data Bases. NY, US, pp: 428-439.

Shimin, S., 2006. Data stream clustering method based on grid and density [D]. Ph.D. Thesis, Dalian University of Technology, Dalian.

Wang, W., J. Yang and R. Muntz, 1997. STING: A statistical information grid approach to spatial data mining. Proceeding of the 23rd International Conference on Very Large Data Bases. Athens, Greece, pp: 186-195.

Zhu, W.H., 2006. Based on the arbitrary shape of the data stream clustering algorithm [J]. J. Softw., 17(3): 379-387.