Advanced Split BIRCH Algorithm in Reconfigurable Network

Wan Li and Yang Jie
Department of Computer and Communication Engineering, Hunan University of Science and Engineering, Yongzhou, Hunan, 425199 China

Abstract—The Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) has a disadvantage that it reduced the accuracy of the arbitrary shape clustering algorithm clusters, to this end a split improved BIRCH algorithm (AS-Birch) was put forward. Through the analysis of the reconfigurable network and a detailed analysis of application scenarios and functional requirements of business clustering, according to the different characteristics of each clustering algorithm, we choose a hierarchical clustering algorithm BIRCH as a business solution to the clustering programs. On the basis of the drawbacks in BIRCH algorithm, we designed split improved BIRCH algorithm and the specific steps of the algorithm is given. Finally, Advanced Split BIRCH algorithm was applied to a detailed experimental simulation experiments show in reconfigurable network. The experiment suggests: AS-BIRCH algorithm has a greater improvement in terms of packet loss rate, delay and jitter than BIRCH algorithm, especially when dealing with irregular shape cluster, AS-BIRCH algorithm greatly improved the accuracy rate concerning to the business clustering.

Index Terms—Multidimensional Cognitive Synergy; Clustering Algorithm; Intrusion Detection; Artificial Data Sets

I. INTRODUCTION

Cluster analysis is based on the similarity of the data objects to group and found the distribution of the data space, and it is also main method of a class of data mining. In the field of data mining, because there is a very large and complex data sets need to handle, the two needs proposed by traditional clustering methods should meet the requirements: a) dealing with numerical attributes and categorical attributes at the same time; b) the efficiency of the algorithm should meet the requirements of large data sets. BIRCH algorithm using clustering feature (CF) and the clustering feature tree (CF tree) structure, which saves the cost and memory overhead of I/O, and makes its cost and the size detesting linear, which simply scan data set at a time can be produce a higher quality of clustering. Therefore it suitable for large data sets [1-3].

The traditional clustering algorithms can be divided into five categories, namely division method, hierarchical method, density-based method, grid-based methods and model-based methods [4]. These algorithms are just looking for similar objects in a single table, but now a lot of data stored in a relational database. How to use a relational database, mine the knowledge from relational database, and get the clustering results are now important directions in researching clustering [5-7].

BIRCH algorithm is a very effective, traditional hierarchical clustering algorithm, which can be used to effectively clustering scan again, and be able to effectively deal with outliers. However, the algorithm only considers the cluster relationships between each data object, while ignores the relationship between the cluster and the cluster, so that when their size are difference the quality of the cluster results are relatively low. Since the algorithm uses the same threshold T to generate each cluster at the whole time, the clustering results are almost a plurality of clusters having the same diameter [8-10]. Therefore, the algorithm is suitable for the clusters with the similar volume size, and to clusters having difference volume is hard to get satisfied clustering results.

In recent years, the development of information networks gradually moving towards diversification of application requirements, architecture isomerization and security control management. The existing network system is still dependent on IP-core non-connected group to achieve inter-connected transmission of information, which leads to a series of problems, including a rigid network structure, the lack of intelligence in core functions, and so on. So that the existing network system is now very difficult to support the diversified network services and high-credible information services [11-13]. Currently some research achievements dealing with the original Internet did not get rid of the status of tight coupling of user services and network services [14]. After the emerging Businesses’ the type and range of applications expanded to a certain limit in the future, the transformation of the network will face increasing more difficulties. Therefore, the development of new generation network architecture has become a top priority. In addition, some of the country’s research institutions trying to overthrow the old architecture, proposed a series of new network architectures. United States’s FIND project was launched in 2005; it put forward solutions including the switching and recursive network architecture, the user to controllable routing mechanism and network virtualization. The 4WARD projects of the European Union have the scalable virtual network service delivery capacity. Japan the AKARI project, by means of
network virtualization, support diverse data access and network extensions [15-16].

Section 2 of this paper describes the basic concepts of the reconfigurable network and business clustering; Section 3 analyzes the characteristics of various types of clustering algorithms to select the most suitable for business clustering algorithm based on Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) algorithm; Section 4 proposed the Advanced Split BIRCH (AS-BIRCH) algorithm; Section 5 through experiments verified the effectiveness of the algorithm.

This paper mainly made to expand and innovative work in the following areas:

On the basis of detailed analysis of business clustering application scenarios and functional requirements, according to the different characteristics of each clustering algorithm, select the hierarchical clustering algorithm BIRCH as business clustering solutions. For Balanced Iterative Reducing and Clustering using Hierarchies, i.e. BIRCH algorithm has a disadvantage that it reduced the accuracy of the arbitrary shape clustering algorithm clusters, to this end a split improved BIRCH algorithm (AS-Birch) was put forward. The specific steps of the algorithm is given.

In order to verify the correctness and validity of the split improved BIRCH algorithm (AS-BIRCH), we used the split improved BIRCH algorithm to deal with business cluster in the detailed experimental simulation of the reconfigurable network. Simulation results analysis suggests: AS-BIRCH algorithm comparing with BIRCH algorithm, has made progress in dealing with irregular shape clusters.

II. RECONFIGURABLE NETWORK

In order to get rid of the shortcomings of traditional network like curing functions, single structure, its difficulty in re-planning deployments and upgrades, reconfigurable network system put forward the Business - Meta-Service - Meta-Ability model. In this model, the distinguishing of the network business cluster is an important prerequisite. The business clustering module deployment scheme is shown in Figure 1.

When the Reconfiguration Goal (RG) System’s needs changed, each functional entity of the state wants to enter through the reconstruction. The reconfigurable target consists of two parts:

(a) Goal Time (GT). The time limit for achieving functional goals, i.e. It must be completed within the number of time units. Its relationship with the the Reconfigurable Goal (RG):

\[ GT \cap Goal \Rightarrow (GT \rightarrow RG) \]  

(b) Goal Resource (GR). In the process of achieving Reconfiguration Goals, each functional entity will consume some Goal Resource, such as CPU, memory, storage, etc. But the resources are limited, so the achievement of Reconfiguration Goals can not be to occupy or consume too many resources. Its relationship with the Reconfigurable Goal (RG):

\[ GR \cap Goal \Rightarrow (GR \rightarrow RG) \]  

The formal definition of the network business clustering:

\[ C = \{c_1, c_2, ..., c_n\} = f(R) \]  

In which \( R = \{r_1, r_2, ..., r_k\} \) represents business collection in reconfigurable network, \( r_i \) represents a certain type of QoS business needs, \( f \) represents clustering algorithm, \( c_i \) represents a kind of business-class clusters. They all will meet the following criteria:

\[ \cup_{j}^{i} r_i \cap T; (2)c_j = \{t, \forall t \in c_i, t \rightarrow s_j\} \]  

\[ \cap_{j}^{i} c_j = T; (2)c_j = \{t, \forall t \in c_i, t \rightarrow s_j\} \]  

\[ \cup_{j}^{n} S_j \cap \cup_{j}^{n} r_j \cap \cup_{j}^{n} \rightarrow 0 \]  

\[ \cup_{j}^{n} S_j \cap \cup_{j}^{n} r_j \cap \cup_{j}^{n} \rightarrow 0 \]  

In which \( S_j \) represents the set of Meta-Service. It can be seen that the objectives of the network service clustering is the different divisions of the business in accordance with the demand for multiple clusters, which makes the reconfigurable network able to provide the most appropriate set of Meta-Service, and achieves the mapping of network business and the Meta-Service.

III. PROPOSED SCHEME

Common clustering algorithm can be divided into separate clustering, density-based clustering, hierarchical clustering, and so on. In dealing with the real condition of the reconfigurable network, the business clustering algorithm should satisfy the following two constraints:

**Constraints 1:** Due to a variety of new business that may arise in the future, network management can not specify the number of clusters in advance for clustering algorithm.

**Constraints 2:** in response to the rapid increase of the number and types of business, the algorithm should be capable of handling large-scale data. This puts
requirements on the algorithm’s algorithm complexity and space complexity.

As can be seen, because of the presence of constraint 1, so that the algorithm like k-means need to be specified the number of clusters in advance can not be applied to the business clustering. While constraint 2 limit the clustering algorithm like DBSCAN which needs great time and space. So this paper will select the hierarchical clustering algorithm for business clustering.

BIRCH is a typical hierarchical clustering algorithm. For business clustering, the algorithm has the following advantages:

(a) The algorithm does not record the contents of each data object, but only to maintain the summary information for each cluster, significantly saved storage space.
(b) The storage space being occupied by the algorithm can be specified by the user. The storage space does not increase with the increase of the input data.
(c) The algorithm can produce different levels of output according to the requirements of the network management, which means that the management level can adjust the cluster size and quantity, increasing the flexibility of the resources and services over the configuration.
(d) The algorithm is an incremental algorithm, in which the new business data can be dynamically updated for clustering.

To sum up, the BIRCH algorithm can relatively meet the various needs of the business clustering, but the algorithm also has some shortcomings. The next part of this paper will discuss how to improve BIRCH algorithm and how to apply it to business clustering.

<table>
<thead>
<tr>
<th>Parameter Description</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Branching factor B</td>
<td>Non-leaf node in storing the number of CF</td>
</tr>
<tr>
<td>Threshold T</td>
<td>Each leaf node represents the radius of the cluster</td>
</tr>
<tr>
<td>Leaf node capacity L</td>
<td>The he number of CF node can store</td>
</tr>
<tr>
<td>Memory capacity M</td>
<td>The limit of the whole CF tree takes share of intrinsic maximum capacity</td>
</tr>
</tbody>
</table>

### A. Algorithm Based on Split Optimized Birch

BIRCH has advantages including quick processing speed, capable of handling large-scale data. But due to the mechanism, BIRCH algorithm is prone to large errors in the cluster of arbitrary shaped clusters.

This paper analyzes the reasons why BIRCH algorithm exist such defects, and put forward AS-Birch based on BIRCH.

### B. Birch Algorithm

BIRCH algorithm is a divisive hierarchical algorithm. It’s biggest feature is that it is suitable for handling large-scale data. Nowadays, algorithm in network traffic analysis, intrusion detection and other fields has achieved good application. The Clustering Feature (CF) tree is the most critical step in BIRCH algorithm.

The basic structure of CF tree is clustering characteristics of the CF. The structure recorded the number of data points, and all the data points of the linear and square of all data points in the cluster.

![Figure 2. The structure of the CF tree](image)

#### C. Algorithm Defects

During clustering, BIRCH algorithm is based on the Euclidean distance of each data point and each centroid CF BIRCH algorithm to determine which cluster the point belongs to.

A node is stored in the CF tree, when CF number is bigger than the threshold, the node has to split. The splitting strategy is that: firstly, selected the two farthest centroid CF to create two new nodes, and then remaining CF in the original node be allocated for each CF should in accordance with its distance from the new node. Every CF will be assigned to the two nearest new nodes. Lastly, to replace the old node with the new two nodes. If the split is the root node, you need to create a new root node as the parent node of the two split nodes.

This split rule’s disadvantage is that it reduced the accuracy of the clustering algorithm clusters of arbitrary shape.

As shown in Figure 3, each point represents a CF, and these CFs originally belong to the same node. These nodes will now split into two nodes N1 and N2. During the split process, first elect two furthest CF, i.e. black dots in Table 3 P1 (0,2) and P2 (6,2). If allocate remaining CFs in accordance with the distance between p1 and p2 to N1 and N2, the CFs’ abscissa which is less than 3 will all be allocated to N2, which led to two CFs which similar to N1 been assigned N2, and so that the area of N2 represented by clustering is expanded, the cluster centers of N1 and N2 also appear in certain deviation.

Because the judgmental accordance of the classification of a new data object is based on the distance between this object and sub-cluster center, so deviation defect caused by the splitting algorithm will further affect the subsequent clustering results.

![Figure 3. Wrong split](image)
D. The Advancement on BIRCH Algorithm

In real business clustering process, even make a standardized process for the business data in advance, it is difficult to ensure that the natural shape of the business clusters is spherical.

Based on the former analysis we can see, the problem of irregular shape of the clusters will reduce the algorithm clustering accuracy rate, thus affecting the final results of the business clustering. To solve this problem, we designed the AS-BIRCH algorithm, the algorithm uses the idea of density split algorithm to improve BIRCH, making the advanced split algorithm can come closer to the natural shape of the clustering results.

AS-BIRCH also split first by selecting the farthest two CFs, and use these two CFs to generate a new node. When allocate the remaining CFs, AS-BIRCH no longer allocation CFs according to the Euclidean distance between \( CF_1 \) and \( CF_2 \), but according to its minimum nodedistance. The specific definition is as follows:

**Definition1.** The minimum nodedistance nodes \( D(CF_i, node_j) \) will calculated according to equation (7):

\[
D(CF_i, node_j) = \min_{CF, node} \text{dist}(CF_i, CF_j) \quad (7)
\]

\( \text{dist}(CF_i, CF_j) \) means the Euclidean distance of the center of the two CFs.

**Definition2.** Farthest node \( R(CF_i, node_j) \) distance can be calculated from the formula (7):

\[
R(CF_i, node_j) = \min_{CF, node} \text{dist}(CF_i, CF_j) \quad (8)
\]

\( \text{dist}(CF_i, CF_j) \) means the Euclidean distance of the center of the two CFs.

AS-BIRCH algorithm split strategy is no longer like BIRCH algorithm, the CFs were allocated one by one according to the Euclidean distance transferred to two new nodes, but two new nodes in accordance with the most recent node distance to absorb the original node CF. Absorption order should in accordance with the most recent node distance from small to large in turn absorb the CFs. The specific algorithm is shown in Table 2.

The Algorithm’s Step 5 to Step 16 is actually the absorption process of the CF in accordance with the density. During each cycle, select the CF which has the nearest D from a new node, and transfer it to a new node. With the new node set of the CF continues to expand, the to-be-distributed CF with two new nodes’ D is changing, every transfer of a CF need to be recalculated. Thus, compare to the original algorithm which is only by distance distance to allocate CF splitting node, AS-BIRCH algorithm uses the idea of the density absorbing the split of the node, ensure that the cluster formed after the split has good the density to reduce the error rate of the clustering of irregular shape cluster.

BIRCH clustering algorithm’s time consumption is mainly concentrated on the two sub-algorithm CF tree split and inserting data into the CF tree. The assumption that the algorithm processed \( n \) data to establish a CF tree of height \( h \), the CF tree’s split sub-algorithm time complexity is \( O(Bh + 1) \), data inserting into sub-algorithm time complexity is \( O(n \cdot B \cdot h) \).

AS-BIRCH algorithm uses the idea of density split, the split sub-algorithm time complexity is \( O(BH + 2) \), introns algorithm complexity unchanged.

Thus, in the data set \( B \) and \( h \) values smaller, the proportion occupied by the time consumed by the split sub-algorithm in the overall processing time is smaller.

AS-BIRCH, may be appropriate to increase the value of the parameter \( T \), thereby reducing the number of nodes of the CF tree and height \( h \), narrowing the BIRCH gap of AS-BIRCH’s run time.

<table>
<thead>
<tr>
<th>TABLE II.</th>
<th>STAGE PROCESS OF SPLIT NOTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>algorithm: SplitNode</td>
<td></td>
</tr>
<tr>
<td>import: The node to be split old._node</td>
<td></td>
</tr>
<tr>
<td>output: new._node1,node2</td>
<td></td>
</tr>
<tr>
<td>(a) ( n \leftarrow ) old._node CF number</td>
<td></td>
</tr>
<tr>
<td>(b) if ( n \leq r ) or ( n \leq L )</td>
<td></td>
</tr>
<tr>
<td>(c) return? null</td>
<td></td>
</tr>
<tr>
<td>(d) from old._node Pick out the most distant pair CF: CF</td>
<td></td>
</tr>
<tr>
<td>(e) for ( i = 1 ) to ( n - 2 )</td>
<td></td>
</tr>
<tr>
<td>(f) for ( i = 1 ) to ( n )</td>
<td></td>
</tr>
<tr>
<td>(g) ( D_i \leftarrow D(CF_i,new._node1) )</td>
<td></td>
</tr>
<tr>
<td>(h) ( D_i \leftarrow D(CF_i,new._node2) )</td>
<td></td>
</tr>
</tbody>
</table>
| (i) if \( (\min(D_i,D_j) < \min_{
\text{dist}}) \)                             |
| (j) \( t \leftarrow c \)                     |
| (k) if \( D_i \approx D_j \)                |
| (l) shift ToNode1 \leftarrow True          |
| (m) else                                   |
| (n) end if                                 |
| (o) end for                                |
| (p) return new._node1,new._node2           |

IV. EXPERIMENTAL RESULTS

In order to verify the the AS-BIRCH algorithm in dealing with clusters of irregular shape’s progress than BIRCH algorithm, in this section we constructed irregular shape of the artificial data sets, testing the cluster’s precision on the data set.

Then according to the business classification standard defined by the DiffServ, randomly generated data set multi-type business needs, and then use AS-BIRCH algorithm to the data set clustering. Thus verify the effects of the test algorithm in the business clustering. The experiments is based on a PC which has deployment of Pentium 4 CPU, 3.2 GHz, 1 G memory, and the algorithm implemented by the java language.

A. Artificial Data Set Test

The experiment first made three artificial configured dataset contains clusters of irregular shape: DS1, DS2, DS3 (Figure 4), to compare the effect of AS-BIRCH and BIRCH in clustering of irregular shape clusters. DS1 and
DS2 contains 1600 data points, the DS3 contains 800 data points.

![Figure 4. Test data set](image)

In the present experiment, the output mode of the clustering results is: to select CF tree’s $h_{res}$ level node as the final clustering output results, that is, the layer of every node represents a cluster; CF centroid $h_i$ layer node to describe the shape of the cluster, the $h_{res}$ clustering results marked as the $h_i$ node. The present experiment is set as $h_{res} = 2, h_i = 4$.

Parameters $T$ decides the size of the CF tree. In this study, the main observation of the CF tree is the build results, therefore in order to avoid the memory occupied by the CF tree exceeding the threshold value $M$ and causing the compression and reconstruction of the tree. The experiment will be the first using incremental approach to determine $T’s$ size;

The value of $T$ in the three experiments were 0.2, 0.17, 0.07 separately. Parameter $M$ is set to the maximum amount of memory that can be occupied by CF tree of the height of $h_s$. B and L were taken 6 and 10.

Figure 5 is the DS1 test results. In the BIRCH clustering algorithm results, the right side of the left oval was misclassified to the right oval clusters; while in the AS-Birch of clustering results, only a tiny percentage point being classified error. DS2 test results was shown in Figure 6, the AS-Birch algorithm has more obvious advantages in the test.

The results of tests conducted on the DS3 was shown in Figure 7, the AS-Birch’s clustering error rate is far less than BIRCH. It can be seen that in dealing with clusters of irregular shape, the AS-Birch of clustering correct rate has improved greatly compare with BIRCH.

![Figure 5. The DS1 Test Results](image)

![Figure 6. The DS2 Test Results](image)

B. Business Clustering Simulation

The experimental data set consists of the simulatively generated network business QoS needs. Because the category is different, QoS needs of each business also vary.

In our study, we choose packet loss rate, delay and jitter of three indicators as the QoS demand characteristics of the business, and the business cluster.

Among the simulation process of the generation of network business, the demand characteristics of the business will according to the definitions of DiffServ to generate the four business sets. Defined each business set the range of packet loss rate, delay and jitter.

The definition were shown in table 3.

A set of data sets will simulate 4000 DiffServ defined network services, where each business collection contains 1000 randomly generated business. The data sets for each business are recorded in the collection of the business to which it belongs, used for clustering results for evaluation.

The experiment firstly set and generate testing data sets according to Table 3, and call the AS-Birch and BIRCH cluster test set. Before calling the algorithm data set needs normalized processing, simultaneously randomly disrupt the order of data input. Algorithm
parameter settings: threshold $T = 0.01$, the branching factor $B = 6$, leaf node capacity $L = 10$.

In the test set the correct clustering business share ratio as an evaluation criterion, the results of the experiment are shown in Figure 8. As can be seen from the results in business clustering AS-BIRCH algorithm clustering accuracy rate is higher than BIRCH.

V. CONCLUSION

This paper studies the reconfigurable network business clustering problems. On the basis of detailed analysis of business clustering scenarios and functional requirements, according to the different characteristics of each clustering algorithm, this paper selected the hierarchical clustering algorithm BIRCH as a business clustering solutions; For the BIRCH’s own shortcomings, an advanced algorithm AS-Birch, has a better adaptability in dealing with clusters of arbitrary shape on this issue. Lastly, we verified by experiments the AS-BIRCH algorithm has a good effect in dealing with the network business. Although AS-BIRCH dealing with clusters of arbitrary shape has been greatly improved, but in the experiment we found the input sequence data will still have a certain impact on the clustering results, so the algorithm needs to be improve in this regard. In addition, future studies are necessary for data collection and in-depth analysis on the QoS requirements of the various types of business in the real network environment, so as to lay the foundation for further experiments in the future.

ACKNOWLEDGMENT

The paper is supported by key discipline for computer application and technology of Hunan University of Science and Engineering. This work was supported in part by a grant from the Research Foundation of Education Bureau of Hunan Province, China (Grant No.12C0681)

REFERENCES


Wan Li was born in Hubei China in May 9, 1983. Graduated from London South Bank University in 2007 in U.K., and was granted a Master’s Degree of Science in Internet and Multimedia Engineering. Previous education was received in Central China Normal University, and was granted a Bachelor’s Degree of Science in Electronic and Information Engineering. Major field of study is intelligent computing, image processing and pattern recognition.

He works in Department of Computer and Communication Engineering, Hunan University of Science and Engineering as lecturer.

Mr. Wan is a membership in key discipline for computer application and technology of Hunan University of Science and Engineering.

Yang Jie was born in Hunan China in November 23, 1976. Graduated from Central South University in 2008 in China, and was granted a Master’s Degree of Engineering in Computer Application Technology. Previous education was received in Northwest University for Nationalities, and was granted a Bachelor’s Degree of Engineering in Computer Application Technology. Major field of study is network security, image processing, artificial intelligence.

He works in Department of Computer and Communication Engineering, Hunan University of Science and Engineering as associate professor.

Mr. Yang is a membership in key discipline for computer application and technology of Hunan University of Science and Engineering.