Automation, Control and Intelligent Systems
2015; 3(5): 71-75
Published online September 28, 2015 (http://www.sciencepublishinggroup.com/j/acis)
doi: 10.11648/j.acis.20150305.12
ISSN: 2328-5583 (Print); ISSN: 2328-5591 (Online)

Electric Power Remote Monitor Anomaly Detection with a Density-Based Data Stream Clustering Algorithm

Liyue Chen[1], Tao Tao[1], Lizhong Zhang[1], Bing Lu[1], Zhongling Hang[2]

[1]Electric power dispatching control center, State Grid Zhejiang Electric Power Company, Hangzhou, China
[2]Department of Automation, Shanghai Jiao Tong University, Shanghai, China

Email address:
chenly@zj.sgcc.com.cn (L. Chen), taotao980925@qq.com (T. Tao), zlz951@163.com (L. Zhang), lubing007@126.com (B. Lu), daba@sjtu.edu.cn (Z. Hang)

To cite this article:
Liyue Chen, Tao Tao, Lizhong Zhang, Bing Lu, Zhongling Hang. Electric Power Remote Monitor Anomaly Detection with a Density-Based Data Stream Clustering Algorithm. *Automation, Control and Intelligent Systems*. Vol. 3, No. 5, 2015, pp. 71-75.
doi: 10.11648/j.acis.20150305.12

Abstract: Nowadays data streams are more and more involved in the real industry. In this paper, the authors apply the data stream clustering to the electric power remote anomaly detection and propose a new data stream clustering algorithm based on density and grid (density-based data stream clustering algorithm, DBClustream). The double-frame analysis model is used in the proposed algorithm. In the online component, the authors optimize the initialization of the parameters for the K-means algorithm with a method based on density and grid and use the kernels to represent the micro clusters as the result of the online component. In the offline component, the time fading weight and the dynamic threshold to optimize the performance of the DENCLUE algorithm are proposed. To evaluate the performance of the proposed algorithm, both the evaluation of the anomaly detection and the evaluation of the data stream clustering are adopted. As the experiment result demonstrates, compared with the others algorithms, DBClustream can resolve the multi-density data stream and keep the high detection rate as well as the low false positive rate.

Keywords: Anomaly Detection, Density Based, Clustering Algorithm, Data Stream

1. Introduction

Because of the rapid development of the smart grid, more and more unmanned substations with automation equipment and telecommunication systems come out. And the requirement of the safe daily operations for the centralized monitoring of the substations becomes increasingly high. Any problem in the automated remote system might put the substation in danger and lead to a large accident in the power grid [1]. Currently, the electric remote monitoring relies on prior knowledge and post-hoc analysis, which is hysteretic and cannot satisfy the increasingly complex needs of the power grid.

By monitoring and analyzing the information of the communication between the substations and the control center, the anomalies and the problems can be found as soon as possible, which brings great convenience for the latter exception handling and reduces the possible economic loss. However, a huge amount of the data and the unknown anomalies raise the high requirement to the anomaly detection methods. Compared with other industry fields, the data of the electric power remote monitor system is mainly characterized by the following features [2] [3]:

1) Discrete: the data acquiring devices periodically collect the information of the substations and send it to the control center;
2) Non-spherical distribution: the data of remote monitor system presents an irregular non-spherical distribution with a number of dense centers;
3) Complexity: the complexity of anomaly detection is proportional to the number of substations and the length of the time to analyze.

Therefore, electric power remote monitor anomaly detection is very difficult.

In this paper, a density-based data stream clustering algorithm (DBClustream) is proposed. It adopts CluStream clustering framework, which has two components. In the online component, DBClustream uses the density-based improved K-means to form the micro-clusters, and uses the density-based method to optimize the initialization of the parameters. In the offline component, DBClustream applies the improved DENCLUE algorithm with fading window model to the core points of the micro-clusters. With the windows model and the dynamic threshold, the proposed algorithm has the abilities to analyze the multi density data stream with the non-spherical distribution.

The paper is organized as follows. In section 2, the related work is briefly discussed. In section 3, the basic conceptions and definitions are presented. In section 4, the proposed algorithm is explained in details. The experimental results of the proposed algorithm are shown in the section 5. In the end, the section 6 is the conclusion of the paper with some directions of future works.

2. Related Works

In the recent twenty years, anomaly detection has become a hot topic in many industries, such as network security and production process. And the reliable testing standards and the viable methods are summed up. Among the viable methods, clustering algorithm is one of the most popular algorithms that have received attention in many fields. The typical clustering algorithms can be divided into the following categories: partitioning method, hierarchy method, density-based method, grid-based method and model-based method.

K-means clustering algorithm is the most frequently used partitioning algorithm. All the points in the dataset are assigned to *k* groups with *k* cluster centers. DBSCAN (Density-Based Spatial Clustering of Application with Noise) algorithm and DENCLUE (DENsity-based CLUstEring) algorithm represent two main directions of density-based method. The basic idea of DBSCAN is that for the point in a cluster, the points counted in its neighbor can't be less than a minimum of user setting. [4] As the improved DBSCAN algorithm, IDBSCAN [5] cuts down the execution time and LD-BSCA [6] reduces the number of the input parameters. And the basic idea of DENCLUE is to model the overall point density analytically as the sum of influence functions of the data points. CLIQUE (CLustering In QUEst) algorithm is a combination of the density-based method and the grid-based method. The basic idea of CLIQUE is to distinguish the sparse area from the crowded area in the data space and to form the clusters with the crowded area.

During the research on the data stream, Aggarwal proposed CluStream clustering algorithm, which is a double-frame analysis model and includes online and offline stream processing. [7] The online part uses K-means algorithm to form a number of micro clusters and the offline part uses the improved K-means algorithm to realize the macro clustering and cluster evolution. The double-frame analysis model is adopted in the later data stream algorithm research.

Among existing data stream clustering algorithms, Den-Stream [8], DDenStream [9], D-Stream [10] and MR-Stream are algorithms based on density based clustering. All of them can detect arbitrary shape clusters as well as handling noise. However, the quality of these algorithms is decreased in multi-density data where different regions have various densities [11]. All the algorithms focus either on the quality of the clustering or the anomaly detection and barely consider both.

3. Basic Conceptions

Definition 1: *Data point weight:* The initial weight of the data point is 1. And the weight of the data point *x* in the time t_n is defined based on the weight in the time t_0 :

$$w(x,t_n) = w(x,t_0)\times f(t_n - t_0) \tag{1}$$

The function *f* is the fading function.

Definition 2: *Time fading function:* The weight of data points or micro clusters is decreased exponentially over time via the time fading function as follows:

$$f(t_n - t_0) = \begin{Bmatrix} 2^{-\lambda(t_n - t_0)}, 0 < t_n < (t_0 + t_{threshold}) \\ 0, t_n > (t_0 + t_{threshold}) \end{Bmatrix} (\lambda > 0) \tag{2}$$

Definition 3: *Grid weight:* For a grid at the time t_n , the grid g_i weight is defined as the sum of the weight of the data points that are in the grid:

$$w(g_i, t_n) = \sum\nolimits_{x\in g_i} w(x,t_n) \tag{3}$$

Definition 4: *Dense grid:* At the time t_n , when the grids are sorted by the grid weight, the grids that are in the top one-tenth are dense grids.

Definition 5: *Kernel weight:* For a kernel at the time t_n , the weight of the kernel k_i is defined as the sum of the weight of the data points in the cluster c_i hat the kernel represents:

$$w(k_i, t_n) = \sum\nolimits_{x\in C_i} w(x,t_n) \tag{4}$$

4. Density-Based Data Stream Clustering Algorithm

DBClustream has an online component and an offline component. In the online component, DBClustream uses the improved K-means algorithm to get the micro-clusters. The initial cluster centers are chosen based on grid and density, making the algorithm more stable and more accurate. In the offline component, the micro-clusters are represented by their core points. DBClustream generates the final clusters with the improved DENCLUE algorithm.

4.1. Online Phase of DBClustream

Table 1. *Notations Used In Online Phase.*

n	The total number of data points
m	The total number of grids
m_d	The total number of dense grids
k	The total number of initial cluster centers
l	The number of selected features
ε	The threshold on the change of the cluster centers

In this section, we discuss how to use the density-based improved K-means to form the micro-clusters. The used symbols are listed in Table 1.

As the definitions, the relations of parameters are as follows:

$$m_d = m/10 \tag{5}$$

$$k = m_d \tag{6}$$

As the result of the online phase, the micro-clusters should be able to minimize the computational complexity of the offline phase under the basic premise that the diversity of the data points is retained. So we set the value of *k* as follows:

$$k \approx \sqrt{4 \times 1 \times \frac{n}{10}} \tag{7}$$

With (5) and (6), we can get the initial value of the parameter *m*.

The improved K-means algorithm has two steps. In the first step, the data points are mapped into *m* grids. We take the core points of the m_d dense grids as the *k* initial cluster centers. In the second step, with the *k* initial cluster centers, the K-means algorithm is adopted to get the micro-clusters. The pseudo code of the density-based improved K-means algorithm is shown in Table 2.

Table 2. *Pseudo Code of Improved K-means Algorithm.*

```
Improved K-means( DS, ε )
Input: data stream
Output: micro clusters MCs
Calculate the number of the micro clusters k with (7)
The initial kernels of the micro clusters is K[]
K[] = GetInitial Kernels (DS, k)
Do
    $\triangle d = 0$
    while not end of stream do
        Read data point x from Data Stream
        Calculate the distance from x to all the points in K[]
        MCs (K[i]) = MCs (K[i])∪x   (K[i] is the nearest kernel from x)
    end while
    for micro cluster mc in MCs do
        Calculate the new kernel of mc
        $\triangle d +=$ distance between new and old kernel of mc
        Replace the old kernel with the new one
    end for
Until $\triangle d < \varepsilon$
return MCs

GetInitial Kernels (DS, k)
Get the maximum and the minimum value for each selected features in the data space
Divide the data space uniformly into m grids
Map the data points into the grids
Sort the grids by the number of their data points
Get the dense grids and their center points
return the center points of the dense grids
```

4.2. Offline Phase of DBClustream

In the offline phase, the improved DENCLUE algorithm on the micro-clusters is applied to get the final result. DENCLUE algorithm is mainly based on the following ideas:

1. the density influence of each data point to the other points in its neighborhood is descripted by a mathematical function;
2. the global density of the data space can be modeled as the sum of the influence of all data points;
3. the final clusters are obtained by determining the local maximum of the global density of the data space.

In the improved DENCLUE algorithm, the following three changes are proposed:

1. the weight of the kernel is considered in the density influence of the kernel;
2. as none of the kernel is noise point, there is no threshold to eliminate the noise point;
3. a dynamic threshold is applied to merge two adjacent density attractors.

The pseudo code of the improved DENCLUE algorithm is shown in Table 3.

Definition 7: *Kernel density function:* At the time t_n, the density function of the kernel x_i is modeled by the Gaussian influence function and related to the weight of the cluster *c* that the kernel x_i represents.

$$d(x, x_i, t_n) = w(c, t_n) * e^{-\frac{(x-x_i)^2}{2h^2}} \tag{8}$$

Definition 8: *Global density function:* At the time t_n, the global density function is defined as the sum of the kernels density functions. Given *k* kernels described by the vectors $D = \{x_1, \ldots x_k\}$, the global density function is as follows:

$$d(x, t_n) = \sum_{i=1}^{N} d(x, x_i, t_n) \tag{9}$$

Definition 9: *Density attractor:* A density attractor is defined as the local maximum of the global density function.

Definition 10: *Dynamic threshold:* The dynamic threshold is the minimum density for two adjacent density attractors to merge. If we use a global threshold in the data space, the low density clusters cannot be revealed in the high density ones. The dynamic threshold is defined as follows:

$$\xi = \theta \times \min(d(x_1, t), d(x_2, t)) \tag{10}$$

where θ is the dynamic parameter between 1 and 0; x_1 and x_2 are two adjacent density attractors.

Definition 11: *Density reachable:* If the minimum point between two adjacent density attractors is bigger than the dynamic threshold, the two adjacent density attractors are density-reachable and we need to merge the two clusters that the two density attractors represent.

Table 3. *Pseudo Code of Improved DENCLUE Algorithm.*

```
ImprovedDENCLUE (WKs, $\theta$, $\varepsilon$)
    Input: Weighted Kernels WKs
    Output: Clusters C
    Initialize the set of attractor points A as an empty set
    For data point $x$ in WKs do
        $x^* = \text{FindAttractor}(x, \text{WKs}, \varepsilon)$
        $A = A \cup \{x^*\}$
        Add the data point $x$ to the set of points $R(x^*)$ attracted to $x^*$
    end for
    Find all the maximal subsets of attractor points $C \subseteq A$, such that any pair of attractors in C is density-reachable from each other
    for c in C do
        for $x^*$ in c do $C = C \cup R(x*)$ end for
    end for
    return C
FindAttractor(x, WKs, $\varepsilon$)
    $t = 0; x_t = x$
    Do
        Get $B_d(x_t, r)$ as the set of all points in WKs that lie within a l-dimensional ball of radius r centered at $x_t$
        $x_{t+1} = \frac{\sum_{x_i \in B_d(x_t,r)} d(x_t, x_i, t_n) x_i}{\sum_{x_i \in B_d(x_t,r)} d(x_t, x_i, t_n)}; t = t+1$
    until $|x_t - x_{t-1}| < \varepsilon$
    return $x_t$
```

5. Experimental Results

In this paper, the experiment is conducted on a PC with Intel Core Dou i7 2 GHz Processors and 16 GB DDR RAM running Windows 7 operating system. And the DBClustream algorithm is implemented in Matlab. We choose KDD CUP 99 dataset to evaluate the performance of DBClustream. The KDD CUP 99 dataset consists of TCP connection records from nine weeks of LAN net-work traffic by MIT. To assess the clustering quality, we use the most widely used parameter, the cluster purity, which is defined as the average percentage of the dominant class label in each cluster. At the same time, we also adopt the detection rate and the false positive rate to assess the performance of the anomaly detection.

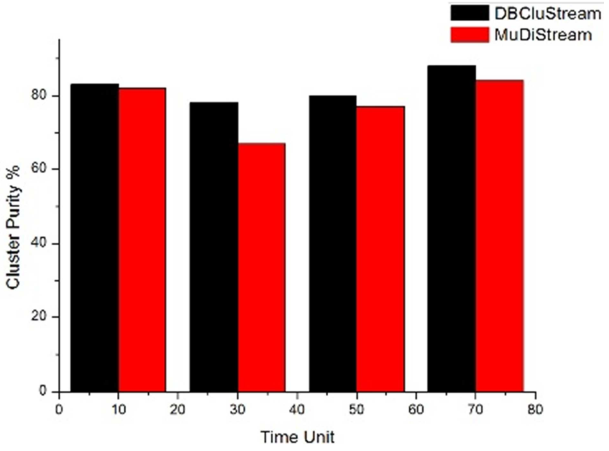

Figure 1. *Comparison of Cluster purity for KDD99 dataset, stream speed = 1000.*

We use the cluster purity to compare the clustering quality of DBClustream and MuDiStream. Figure.1 and Figure.2 show the comparison results of the cluster purity. It can be seen that DBClustream has a very good clustering quality, and is more stable and better than MuDiStream.

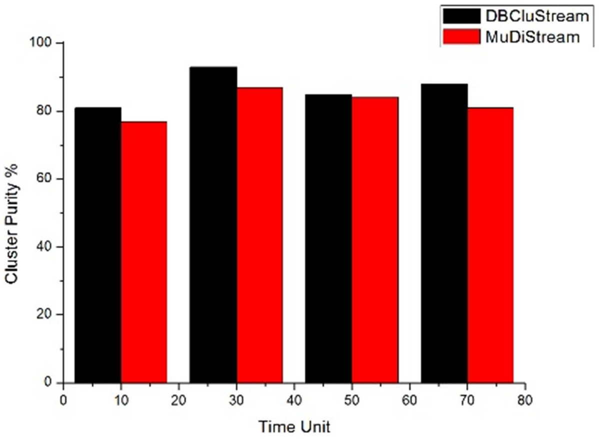

Figure 2. *Comparison of Cluster purity for KDD99 dataset, stream speed = 2000.*

Table 4. *Comparison of detection performance for KDD99 dataset.*

	Detection rate (%)	**False positive rate (%)**
CURE[12]	81.09~85.10	3.47~5.49
Aprior[13]	87.2~87.5	8.1~17.4
DENCLUE	83.2~89.7	5.3~10
DBClustream	85.0~90.2	2.1~4.7

To assess the anomaly detection performance of DBClustream, we compare the detection rate and the false positive rate of DBClustream with other detection algorithms. From Table 4, we can see that compared with the other algorithms, DBClustream keeps the high detection rate and the low false positive rate at the same time. In the electric power remote monitor system, the low false positive rate is as important as the high detection rate, both of which can cut the unnecessary cost for the smart grid.

6. Conclusions

Considering the feature of the electric power remote monitor system, a density-based data stream clustering algorithm (DBClustream) is described for the electric power remote monitor anomaly detection in this paper. The double-frame analysis model is used in DBClustream algorithm. In order to manage multi density data stream, the improved initialization of the parameters of K-means algorithm with a density-based method, the time fading weight and the dynamic threshold methods are applied in DBClustream algorithm. The experimental results show that the proposed algorithm is more effective on the high detection rate and the low false positive rate than other well-known algorithm. In future, we are going to apply proposed algorithm to the real electric power remote monitor system and improve the performance in effectiveness and efficiency.

References

[1] YANG Huan-hong, YE Hai-ming. “Analysis and Monitoring of Electric Power Tele-control Channel Fault,” Journal of Shanghai University of Electric Power, vol. 25, no. 4, pp. 321-324, 2009.

[2] XUE Fei, “The software Design and Implementation of IEC60780-5-104 Protocol,” M.S. the-sis, School of Control and Computer Engineering, North China Electric Power University, Beijing, China, 2012.

[3] WANG Jing, “Research of Online Intelligent Alarm,” M.S. thesis, School of Electrical Sys-tem and Automation, North China Electric Power University, Beijing, China, 2008.

[4] Zhiwei SUN, Zheng ZHAO. “A Fast Clustering Algorithm Based on Grid and Density,” Electrical and Computer Engineering, pp.2276-2279, May 2005.

[5] B. Borah, D. K. Bhattacharyya. “An Improved Sampling-Based DBSCAN for Large Spatial Databases,” Int. Conf. on Intelligent Sensing, pp. 92–96, 2004.

[6] G. Wei, H. Wu. “LD-BSCA: A Local Density Based Spatial Clustering Algorithm,” in IEEE Symposium on Computational Intelligence and Data Mining. IEEE Computer Society, 1999, pp. 291–298.

[7] Aggarwal C, Han J. “A Framework for Clustering Evolving Data Streams,” Proceedings of the 29th VLDB Conference, pp. 81-92, 2003.

[8] F. CAO, M. ESTER. “Density-based clustering over an evolving data stream with noise,” Proceedings of the 2006 SIAM International Conference on Data Mining, pp.328-339, 2006.

[9] M. Kumar, A. Sharma. “Mining of Data Stream Using DDen Stream Clustering Algorithm,” IEEE International Conference in MOOC, pp. 315-320, 2013, doi: 10.1109/MITE.2013.6756357.

[10] Y. Chen, L. Tu. “Density-based clustering for real-time stream data,” Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 133–142, 2007.

[11] Amineh Amini, Hadi Saboohi. “A Multi Density-based Clustering Algorithm for Data Stream with Noise,” IEEE on Data Mining Workshops, pp. 1105-1112, 2013, doi: 10.1109/ICDMW.2013.170.

[12] ZHOU Ya-jian, XU Chen. “Unsupervised Anomaly Detection Method Based on Improved CURE Clustering Algorithm,” Journal on Communications, vol. 31, no. 7, pp. 18-23, 2010.

[13] CUI Guan-xun, LI Liang. “Research on an Intrusion Detection System Based on the Im-proved Apriori Algorithm,” Computer Engineering & Science, vol. 33, no. 4, pp. 40-44, 2011.