A Fuzzy Clustering Based Anomaly Node Detection Method for Publish/Subscribe Distributed Systems

Defeng Wang, Zhuowei Shen* and Wenjun Wu
School of Cyber Science and Engineering, Southeast University, Nanjing, China

*Corresponding author email: zwshen@seu.edu.cn

Abstract. Timely and accurate understanding of the node running status can effectively control the propagation of faults in publish/subscribe distributed systems, which is of great significance to ensure the reliable operation of applications. A method of anomaly node detection based on fuzzy k-means clustering algorithm is proposed. Compared with the traditional K-means algorithm, this algorithm introduces fuzzy membership matrix to realize fuzzy clustering, and uses the idea of local reachability density to improve the selection method of cluster centers. Experimental results show that this method can effectively detect anomaly node in publish/subscribe distributed systems with higher accuracy, recall and F-measure than traditional K-means algorithm. The precision of publish anomaly detection and network anomaly detection is improved by 10.53% and 38.6% respectively.

1. Introduction

Data distribution service (DDS) is a real-time distributed communication middleware specification proposed by OMG[1]. As the most popular publish/subscribe middleware technology at present, DDS adopts publish/subscribe architecture and can ensure real-time, efficient and flexible data distribution. Publish/subscribe distributed systems based on DDS have been widely used in military industry, industrial automation, automatic driving and other critical applications. This kind of applications has higher requirements for the safety and reliability of the system.

In a distributed system, due to fault propagation, an anomaly node will gradually lead to the exception of the whole system. So timely anomaly node detection can effectively control the propagation of system faults, which is critical to ensure the application safety and reliability.

There are usually some normal running patterns in the process of system running. Therefore, by monitoring the system running, collecting and analyze data, the abnormal state of components can be distinguished from normal running patterns. Clustering is an important technology used to discover data distribution and hidden patterns in data mining[2]. The k-means algorithm has a wide application range so far. For data with numerical attributes, it can well reflect the geometric and statistical significance of clustering[3].

Traditional K-means clustering algorithm has some problems, such as difficult to determine the number of clusters, sensitivity to initial cluster centers, low accuracy rate[4]. A fuzzy k-means clustering algorithm is proposed and applied to the anomaly node detection in publish/subscribe distributed systems. By introducing fuzzy membership matrix, the data is clustered effectively based on membership relationships. By introducing local reachability density[5], the selection method of cluster center is improved to promote the accuracy of data division. The experimental results show that the method can detect the anomaly nodes more accurately and effectively.
2. Anomaly Node Detection Model

2.1. Process of Anomaly Node Detection

The process of anomaly node detection for publish/subscribe distributed systems is shown in figure 1.

As shown in the figure, Data acquisition and preprocessing is the basis of anomaly node detection. The status data of components is collected through operating system API and DDS middleware while system running. Then the collected data are preprocessed by multi-source data fusion and standardization to meet the requirements of anomaly node detection. In model training stage, fuzzy k-means clustering algorithm is adopted for classification training with the pre-processed data, which can realize the classification of different running states of components without prior knowledge. Finally, in abnormality monitoring stage, on the basis of the trained model, the fuzzy membership degree discrimination principle is used to detect the anomaly nodes.

2.2. Index System of Node Status Monitoring

The selection of node running characteristic parameters is related to the accuracy of anomaly node detection, so it is critical to establish an accurate and comprehensive index system for node status monitoring[6]. The pub/sub system pays more attention to the node's pub/sub behavior attributes, such as the content and time of pub/sub data. Moreover, considering that the running of software is closely related to the environment state, CPU utilization, memory consumption and network bandwidth are selected as the performance characteristic parameters of nodes[7]. Table 1 shows the system monitoring index architecture.

| Primary index | Secondary index                   |
|---------------|-----------------------------------|
| Time          | Time of data acquisition           |
|               | Time of pub/sub                   |
| Status        | Content of pub/sub                |
|               | Action (pub/sub)                  |
|               | Topic of pub/sub                  |
| Performance   | CPU utility (%)                   |
|               | Memory available (MB)             |
|               | Network bandwidth (bps)           |
3. Fuzzy K-means Clustering Algorithm

3.1. Related Theory and Definitions

Suppose that the running state dataset of nodes in pub/sub distributed system contains $n$ samples, and each sample can be represented as an $m$-tuple vector, which represents $m$ collected index values described in table i. So the dataset can be described by matrix $X$, as shown in equation (1).

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}$$  \( (1) \)

where $x_{ij}$ means the $j$-th characteristic parameter of the $i$-th sample, $i \in [1, n], j \in [1, m]$.

Due to the dimensional difference between the $m$ characteristic parameters of each sample, Z-score normalization is applied to standardize dataset $X$.

Suppose that there are $k$ clusters in dataset $X$, and the set of the cluster centers is $CC = \{cc_1, cc_2, \ldots cc_k\}$. The fuzzy membership matrix of dataset $X$ is shown in equation (2). Among them, $r_{si}$ is the membership degree of the sample $x_i, i \in [1, n]$ to cluster center $cc_s, s \in [1, k]$.

$$R = \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{k1} & \cdots & r_{kn} \end{bmatrix} = (r_{si})_{k \times n}$$  \( (2) \)

According to the clustering criterion, each column in matrix $R$ is independent of each other and satisfies the constraint conditions: $0 \leq r_{si} \leq 1, \sum_{s=1}^{k} r_{si} = 1$.

The membership degree of sample $x_i$ is shown in equation (3)[8]:

$$r_{si} = \frac{1}{\sum_{k=1}^{k} \frac{d_{ik}^2}{d_{ii}^2}} = \frac{1}{\sum_{k=1}^{k} \frac{|x_i - cc_s|^2}{|x_i - c_i|^2}}$$  \( (3) \)

where $d_{is}$ is the Euclidean distance between sample $x_i$ and cluster center $cc_s$, $b \geq 1$ is Fuzzy weighted factor.

Let $d_{\mu}(p)$ is the distance between sample $p$ and its $\mu$-th nearest neighbor, called $\mu$-distance of sample $p$. If the $\mu$-distance of $p$ is known, the $\mu$-distance neighborhood of $p$ contains each of the sample whose distance from $p$ is absolutely shorter than the $\mu$-distance of $p$, denoted by $N_{\mu}(p)$.

Equation (4) [9] defines the $\mu$-reachability distance of sample $p$ to sample $q$.

$$d_{\mu\rightarrow reach}(p, q) = \max\{d_{\mu}(q), |p - q|\}$$  \( (4) \)

Equation (5) represents the local reachability density of sample $p$.

$$ld_{\mu}(p) = 1/\left(\frac{\sum_{q \in N_{\mu}(p)} d_{\mu\rightarrow reach}(p, q)}{|N_{\mu}(p)|}\right)$$  \( (5) \)

3.2. Algorithm Description

The selection of cluster centers is critical in fuzzy k-means algorithm. In general, the actual class centers are far away from each other and the surrounding nodes are dense. Unlike the traditional k-means algorithm which randomly selects the initial cluster centers[10], we select the samples with high local reachability density and low similarity as the initial cluster centers.

As shown in Figure 2, in each iteration, according to their fuzzy membership degrees with respect to the selected $k$ cluster centers, all the samples are added to their corresponding cluster. Then for each cluster, the local reachable densities of its members are recalculated, and the one with the largest local reachable density is selected as the new cluster center.

When the clusters are no longer changed or the maximum number of iterations is reached, the algorithm is terminated.
The following is the specific process of the proposed improved fuzzy K-means clustering algorithm:

```
func Fuzzy-KM(X, k) {
    Input: running state dataset X, number of clusters k
    Output: set of clusters \( C = \{C_1, C_2, \ldots C_k\} \)
            
    for (each sample \( p \) in X)
        calculate \( lrd(\mu(p)) \) according to equation (5);
    Select \( k \) samples in \( X \) which have large \( lrd(\mu) \) and their \( \mu \)-distance neighborhood is disjoint;  
    Add them to the cluster centers set \( CC \);  //find the initial cluster centers
    for (itor =1; itor \leq Max_Itor; i++)
        Get the fuzzy membership matrix \( R \) according to equation (2) and equation (3);
        for (i =1; i\leq n; i++)
            \{  //add each sample to corresponding cluster according to its fuzzy membership degree.
            \( \max_r \) = \( \max(\max_r 1, \max_r 2, \ldots \max_r k) \);
            for (j=1;j\leq k;j++)
                if (\( r_{ji} = \max_r \)) then
                    \( C_j = C_j \cup \{x_i\}, x_i \in X; \)
                    break;
            \}
        for (each cluster \( C_i \in C \))  //update the cluster centers
            calculate \( lrd(\mu(p)) \) according to equation (5) for all \( p \in C_i \);  
            let \( CC^{new} \) be the sample with the maximum local reachability density in \( C_i \);  
            if (all k cluster centers remain unchanged) then
                return \( C \);
        }
    return \( C \);
}
```

**Figure 2.** Flow chart of fuzzy K-means clustering algorithm.
4. Experiments
In an automatic pilot control system based on DDS, the application components include several sensors, data processing controllers and actuators to control the driving of vehicles. Each sensor is responsible for sensing and detecting the corresponding road and vehicle condition information, and delivering the data to the data processing controllers through DDS middleware. In the experiment, one of the sensors, the speed sensor is monitored while the system runs, and its running state data is collected and formed into the dataset[11].

There are three kinds of running states in the process of the sensor operation: normal state, publish anomaly, and network anomaly. The fuzzy k-means clustering algorithm we proposed and the traditional K-means algorithm are executed on the dataset. It is expected that these three states can be distinguished by clustering results.

The experimental results show that both algorithms can divide the samples into three clusters, which are corresponding to the three running states. Figure 3 shows the comparison among the clustering results of K-means algorithm, fuzzy k-means algorithm and manual annotation classification.

![Figure 3. Comparison of clustering results.](image)

Compared with k-means algorithm, the result of fuzzy k-means algorithm is better. The number of samples in each cluster and the number of samples correctly clustered in each cluster are closer to the results of manual annotation.

Table 2 shows the comparison of the precision, recall and F-measure of both algorithms. Compared with the traditional K-means algorithm, the fuzzy k-means algorithm has a significant improvement in term of all the three metrics. In addition, the algorithm has good robustness. In each running state category detection, the precision is high, and there is no significant difference in recall. It is suitable for the anomaly node detection for pub/sub distributed systems.

**Table 2. Performance comparison between the two algorithms.**

| Algorithm      | Publish anomaly | Network anomaly |
|----------------|-----------------|-----------------|
|                | Precision  | Recall  | F-Measure | Precision  | Recall  | F-Measure |
| K-Means        | 84.0%      | 75.0%   | 79.24%    | 67.46%    | 73.68%  | 70.36%    |
| Fuzzy K-Means  | 92.85%    | 92.85%  | 92.85%    | 93.50%    | 94.73%  | 94.09%    |
| Improvement    | 10.53%    | 23.8%   | 17.17%    | 38.6%     | 28.56%  | 33.72%    |

5. Conclusion
Anomaly node detection is one of the important aspects of health management for publish / subscribe distributed systems. By introducing the fuzzy membership matrix and the concept of local reachable density, an improved K-means algorithm is proposed: fuzzy k-means clustering algorithm. Compared with the traditional K-means algorithm, fuzzy K-means clustering algorithm can get higher Accuracy results, the precision, recall and F-measure metrics are significantly improved. It is proved that the algorithm can be applied to anomaly node detection in publish/subscribe distributed systems, which is helpful to judge the running status of system in time and ensure the reliable operation of the application.
References

[1] Pardo-Castellote G. OMG Data-Distribution Service: architectural overview[C]. International Conference on Distributed Computing Systems Workshops, 2003: 200–206.

[2] KHAN S, YAIRI T. A review on the application of deep learning in system health management[J]. Mechanical Systems and Signal Processing, 2018, 107: 241–265.

[3] Kodinariya T M, Prashant R. Review on determining number of cluster in k-means clustering[J]. International Journal, 2013, 1(6): 90-95.

[4] Dong Pengyu, Wang Hongwei, Chen You. Improved System Cluster Analysis Based on Rough Set and General Euclidean Distance[P]. Algorithms, Computing and Artificial Intelligence, 2019: 82-86.

[5] Linlin Shi, Zhenwei Zhou, Yun Huang. Research and Verification of Fault Diagnosis Method for Avionics System Based on Data Mining[P]. Algorithms, Computing and Artificial Intelligence, 2019: 6-12.

[6] SCHUMANN J, ROZIER K Y, REINBACHER T, et al. Towards real-time, on-board, hardware-supported sensor and software health management for unmanned aerial systems[J]. International Journal of Prognostics and Health Management, 2015: 1-21.

[7] Srivastava A N, Schumann J. The Case for Software Health Management[C]. IEEE Fourth International Conference on Space Mission Challenges for Information Technology. Palo Alto, CA, USA: IEEE, 2011: 3-9.

[8] Cebeci Zeynel. Initialization of Membership Degree Matrix for Fast Convergence of Fuzzy C-Means Clustering[C]. 2018 International Conference on Artificial Intelligence and Data Processing, 2018.

[9] Breunig Markus M, Kriegel Hans-Peter, Ng Raymond T, et al. LOF: Identifying Density-based Local Outliers[C]. ACM SIGMOD International Conference on Management of Data. New York, NY, USA: ACM, 2000: 93–104.

[10] Xiaokang Wang, Huiwen Wang, Dexiang Wu, Yihui Wang, Rui Zhou. A Fuzzy Consensus Clustering Based Undersampling Approach for Class Imbalanced Learning[P]. Algorithms, Computing and Artificial Intelligence, 2019.

[11] CORBELLINI A, MATEOS C, ZUNINO A, et al. Persisting big-data: The NoSQL landscape[J]. Information Systems, 2017, 63: 1–23.