A Data Preprocessing Optimization Method based on Sliding Time Window in Communication Alarm Events

Wengang Chen[1, a], WenWei Chen[2, b], Bing Fan[2, c], Runze Wu[2, d], Yanru Wang[3, e], Junli Fan[3, f]

[1] Jincheng Power Supply Company, Shanxi Electric Power Company, Jincheng 04800, China;

[2] School of Electrical and Electronic Engineering, NCEPU, Beijing102206, China.

[3] Beijing GuoDianTong Network Technology Company Limited, Beijing 100070, China

[a] jcchenwangang@163.com, [b]572451718@qq.com, [c]857894455@qq.com,

[d]81627271@qq.com, [e]wangranru@sgitg.sgcc.com.cn, [f] fanjunli1@sgitg.sgcc.com.cn

Keywords: fault analysis, communication alarm, quality evaluation, alarm sequence partition.

Abstract. In communication fault analysis, mining history alarm data to acquire correlation knowledge has become an inevitable trend. Owing to low efficiency of traditional alarm preprocessing methods, a new preprocessing optimization method is proposed. The quality of alarm sequence partition is evaluated by difference alarm time segment using sliding time window, and the optimal partition based on double constraint and K-average is introduce to extract the alarm transactions by sliding time window so as to acquire better preprocessing results. The performance simulation indicates that this method has higher preprocessing efficiency of alarm sequence and overcomes the problem that the traditional methods are not fully applicable to the alarm sequence, which provides support for the analysis of the follow-up alarm data.

1. Introduction

Big data has motivated relevant fields of experts and scholars to explore new tools to mine the information and knowledge hidden in the big data, and it also provides effective information for decision-making. The concept of data mining was formally put forward in the 11th international joint conference on artificial intelligence in 1989[1], nowadays in the context of big data, data mining should be a kind of common information processing technology, which is widely used. The application of data mining in network fault management focuses on the alarm correlation analysis. In communication fault analysis, multiple alarm is compressed into a few alarm containing more information, so a large number of redundant alarm is filtered out in alarm correlation analysis in order to provide data processing method for fault diagnosis and location [2]. The problem of alarm correlation can be converted to the alarm data of association rule mining.

Association rule mining algorithm of input data format is a transactional database, which is discrete, but the alarm sequence is a time for data, so the reasonable algorithm is applied to the original alarm sequence to form a transactional database [3]. Alarm occurs with a sudden, is not always continuous. The interval among the alarm could be very long, so the sliding time window is unsuitable to transform alarm data, otherwise it will produce a large number of empty window without any alarm data, as well as occupies certain resources, and it has unsatisfactory work efficiency. So it is very important to deal with the alarm sequence before using sliding time window. At present, there are most alarm sequence partition algorithms which the number of time segment must be given in advance, and it leads to an unsatisfactory partition result, so it has a bad influence on extracting alarm transactions. The alarm sequence partition algorithm based on double constraint is proposed [4] which can divide the alarm sequence into a series of relatively centralized and independent time segments, thus the evaluation method of the quality of time segment is proposed based on this method, but the evaluation method does not well in evaluating the partition result of the alarm sequence. This paper proposes a new evaluation method of the quality of time segment,

especially constructs a evaluation function to testify the effectiveness of the algorithm. According to the evaluation function, the partition results of the alarm sequence are evaluated, and the optimal partition is determined to improve the preprocessing efficiency of the alarm sequence and guarantees the efficiency and accuracy of association rule mining.

2. Preprocessing method of alarm sequence

The definition of alarm correlation analysis is that a large number of alarm sequences is mined and analyzed to generate less alarm information that can reflect the cause of device fault. So the analysis of alarm correlation mainly concentrates on extracting alarm correlation rules from historical and real-time alarm data without the specific network topology structure and configuration. Once network topology structure and configuration are changed, new association rules will be automatically produced from the corresponding alarm data, these new rules could help realize the fault diagnosis and orientation. To a great extent, the method based on data mining will alleviate the degree of dependence on expert knowledge and network managers' workload. The preprocessing process of alarm sequence is presented in Fig. 1. First, the original alarm data is transformed to form a transactional database with sliding time window; second, the transactional database is used to mining association rules; finally, the rules are obtained to help the analysis of network fault. The algorithm principle of sliding time window is as follows: It is considered that the alarms in the same window are simultaneity and included in a transaction. The sliding time window could transform the original alarm data into transaction data, the purpose of this method is to extract frequent alarms and not miss alarm correlation. Sliding time window is the common method that processes alarm sequence.

Fig. 1 preprocessing process of alarm sequence

Most of the width of time window is fixed, but alarm sequence is usually abruptness and randomness. The alarm is very frequent within a certain period of time, and the alarm may be less after that certain period of time, in this case, the fixed width of time window may cause that alarm transaction will not be obtained fully when the alarm appears frequently, then sliding time window has low operation efficiency when the alarm is not frequent. In order to extract alarm transactions flexibly and adaptively, the alarm sequence can be divided into some time segments which alarm generates frequently with reasonable alarm sequence partition algorithms, in other words, it can be regarded as cluster analysis of one dimensional time sequence.

3. Quality optimization method of alarm sequence partition

The quality of alarm sequence is directly related to the extraction efficiency of the alarm transaction with the sliding time window method. If the alarm time sequence partition is poor, that means the tightness inside the alarm time period is not high and the separation between the periods is not high. This greatly limits the efficiency and effectiveness of the alarm transaction because the relative correlation between the alarm data is always more intensive in a time period. So the correlation degree between the alarm transactions is low, and it will affect the operation efficiency of the association rules.

3.1 Optimal partition method of alarm sequence.

The validity of alarm sequence partition is the evaluation of the results of the partition in order to determine whether the classification and evaluation of the alarm data is valid and correct. The main content is to compare the results of different classification algorithms and different results of the same algorithm under different numbers of time segments [5]. The optimal partition method of the alarm sequence is shown in Fig. 2. In a reasonable range of k, that is $[k_{\min}, k_{\max}]$, we use the same partition algorithm and different partition numbers k to calculate a series of partition results. And we calculate the value of partition quality evaluation function of each result. Then the best value of evaluation

function corresponds to the best partition number $k*$. The reference articles [4] and [6] have gave different answers to how to determine the optimal partition of alarm sequence.

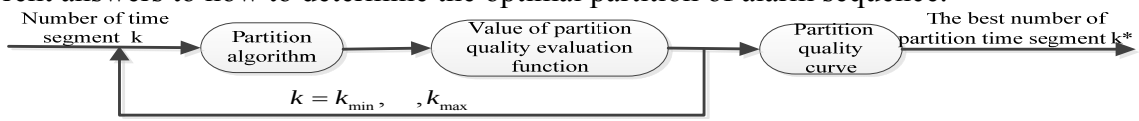


Fig. 2 optimal partition method of the alarm sequence

The evaluation function method based on double constraint in Paper [4] is presented as follows. The nearer alarm distance is better in the same time segment and the further alarm distance is better in different time segment. The distance is the evaluation function of partition time segment. The evaluation function defines two principles. First, each time interval should be compact. Second, the distance between each time interval should be as far as possible. We use $W(t)$ to represent the difference in each time segment and $L(t)$ to represent the difference between each time segment.

For a given alarm sequence $S=(e_1,t_1),(e_2,t_2),\text{K},(e_n,t_n)$, which contains n time points. We will divide the alarm sequence into k time segments. And the quality of the time partition is defined as M. When the value of M is larger , the partition result is better. M can be represented as the ratio of the difference between each time segment $L(t)$ and the difference in each time segment $W(t)$. $W(t)$ is defined as the square of the distance between each time point and the center point of the time segment in the same time segment.

$$M=\frac{L(t)}{W(t)} \tag{1}$$

$$W(t)=\sum_{j=1}^{k}W(t_j)=\sum_{j=1}^{k}\sum_{t\in t_j}d\left(t,\bar{t}_j\right)^2 \tag{2}$$

In(2), k represents the number of the partition time segment. $\bar{t}_j$ represents the center of the time segment j. $L(t)$ is defined as the distance between each center point of the time segments.

$$L(t)=\sum_{1\le l\le j\le k}d\left(\bar{t}_j,\bar{t}_l\right)^2 \tag{3}$$

Paper [6] has proposed a quality evaluation function of K-average. The method constructed a distance cost function and use minimum distance criterion to calculate the optimal time partition segments. The calculation method is as follows.

$$F=L+D \tag{4}$$

In (4), L represents the distance between the time segments. Let L denote the sum of the distance between the center point of each time segment and the center point of all the time points.

$$L=\sum_{j=1}^{k}\left|\bar{t}_j-t_0\right| \tag{5}$$

In (5), t_0 represents the center points of all the data. The meaning of the rest variables is the same in (3). In (4), D represents the distance in the time segment. D is defined as the sum of the distance between each time point and the center point of the time segment.

$$D=\sum_{j=1}^{k}\sum_{t\in t_j}\left|t-\bar{t}_j\right| \tag{6}$$

The meaning of variables in (6) is the same as in (2). When using the distance cost function as the evaluation function of the time partition quality, we use the minimum distance principle. That means when the distance cost function is minimal, the result of the time partition is the best and the optimal choice of the partition time segments k is $\min_k(F)$.

These two methods have different definition of the difference between one time segment and the difference between each time segment, they can't calculate the best number of partition time segment $k*$. According to the characteristics of alarm sequence, we put forward the partition quality evaluation function that is suitable to alarm sequence.

3.2 Evaluation function of partition quality.

An ideal partition should enable the distance between the center points of time segments as far as possible, and the distance between the data and the center of time segment which contains this data is as small as possible. The ratio of the difference between each time segment and the difference of one time segment is greater, the partition result is better. So the best partition is able to maximize the difference between each time segment but minimize the difference of one time segment. Therefore we can construct the evaluation function of partition quality which can be considered to judge a good partition scheme.For a given alarm sequence, it will be formed k time segments that is relatively concentrated and independent, so the evaluation function of partition quality is defined as Q:

$$Q = \frac{I_0(t)}{C_0(t)} \tag{7}$$

In (7), $C_0(t)$ is the difference of one time segment, It is defined as the mean of the average distance between each time point and the center of the time segment which contains that time point.

$$C_0(t) = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{N_i}\sum_{j=1}^{N_i}\left|t_{ij} - \bar{t}_i\right| \tag{8}$$

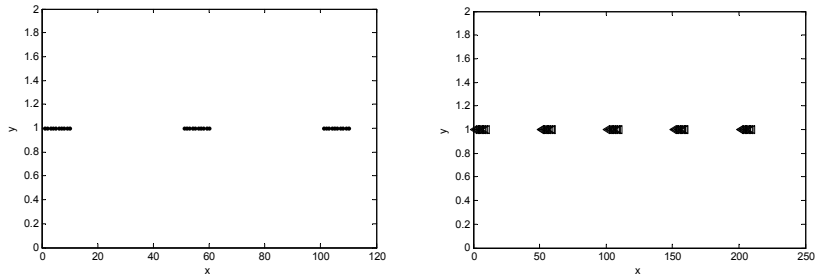
In (8), k is the number of time segments, N_i is the number of alarms in the time segment i, t_{ij} is the time of each alarm, $\bar{t}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} t_{ij}$ is the center data of time segment i. The difference of one time segment is characterized by the cohesion of the time segment. In $C_0(t)$, the distance is based on the Euclidean distance. $C_0(t)$ will be a downward trend if k is increased. $C_0(t) = 0$ when $k = n$, this case is not in conformity with the actual problem, each time point will be one segment and it doesn't make any sense to extract alarm transactions. $I_0(t)$ is the difference between each time segment, it is defined as the average of the distance between the center of the adjacent time segments.

$$I_0(t) = \frac{1}{k-1}\sum_{i=1}^{k-1}\left|\bar{t}_{i+1} - \bar{t}_i\right| \tag{9}$$

The meaning of the variables in (9) is the same in (8). $I_0(t)$ represents the isolation of time segment, the definition of $I_0(t)$ can avoid the repeated computation of distances. Because Q is defined as the ratio of isolation and cohesion, relatively large value of Q corresponds to better partition. The reasonable partition algorithm is adopted to the alarm sequence partition, and we build different partitions under the quality evaluation function, take the maximum value of quality curve in the reasonable parameter range as an optimal partition.

4. Simulations and analysis

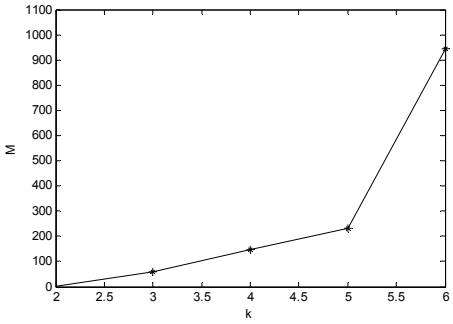
The alarm sequence with 30 and 50 data points which has obvious feature are used to examine the rationality and validity of quality evaluation function introduced in this paper. Experimental data is as shown in Figure 3, the horizontal axis represents the time axis, the vertical represents attribute of the alarm, the attribute is set to 1 when the alarm occurs at a certain time,. It is obvious that the number of time segments is 3 and 5 when the alarm sequence has the optimal partition.



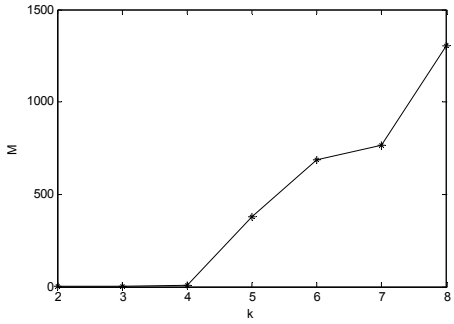
(a) 30 data points (b) 50 data points

Figure 3 Simulating result of experimental data set

The simulating data of Figure 3 is used to the calculation of the evaluation function based on double constraint, M changes with k is as shown in Figure 4. $k \in [k_{\min}, k_{\max}]$, $k_{\min}$ is usually equal to 2, the alarm data are evenly distributed and no obvious characteristic difference when $k_{\min} = 1$. Most scholars use experience rule [13]: $k_{\max} \le \sqrt{n}$, so $k_{\max} = \mathrm{int}\left(\sqrt{n}\right)$. Above all, the scope of k is $[2,6]$ in Figure 3(a), the scope of k is $[2,8]$ in Figure 3(b).



(c) Figure 3(a) data set

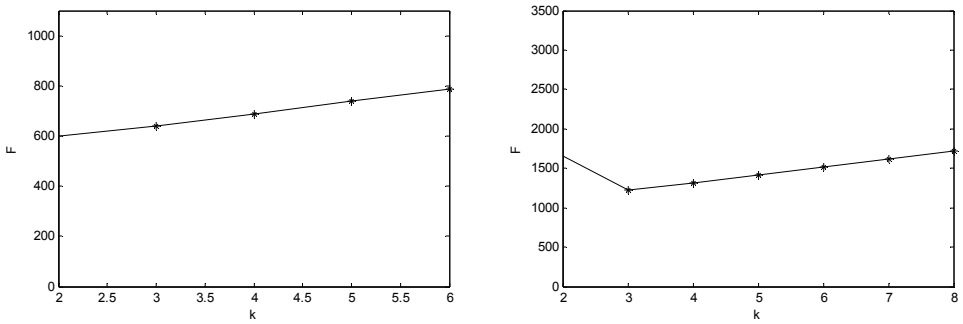


(d) Figure 3(b) data set

Figure 4 the tendency of M based on experimental data set

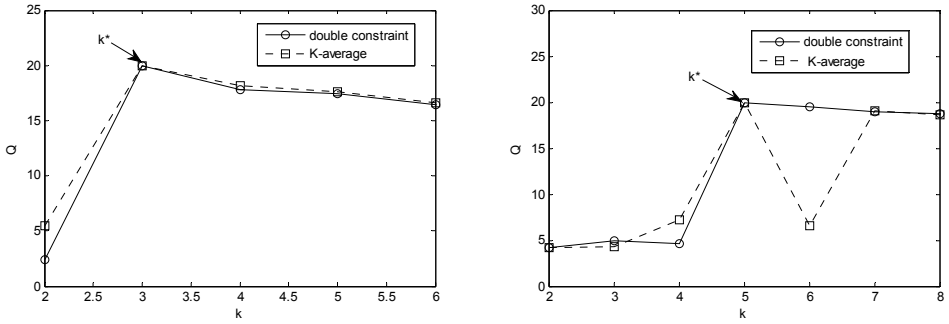
The simulating result of Figure 4 has shown that M is always in a growth trend. M has reached to the maximum at the same time when $k = 6$, it is not in conformity with the obvious characteristics of experimental data set, so the quality evaluation function M for alarm sequence is not very good adaptability. Because M is defined as the difference between the various segments of center distance, considering the distance between each time segment and other segments, the difference between segments is greatly increased, and it will lead to M in a rising trend which is not in conformity with the actual problem. The experimental data of Figure 3 is used to the calculation of the evaluation function based on K-average, F changes with k is as shown in Figure 5. The simulating result of Figure 5 has shown that the trend of F is not stable with the increasing of k. In Figure 5(e), F has been in trend of increasing, F is minimum value when $k = 2$, the optimal partition is two time segments and it do not meet the experimental data set in Figure 3(a) is divided into 3 segments. In Figure 5(f), the change of F increases firstly and reduces lately, then F reaches the minimum when $k = 3$, and it do not meet the characteristics of the experimental data set in Figure 3(b). The division of alarm sequence can be abstracted as a cluster of one-dimensional alarm data, but F is suitable for two-dimensional and high dimensional data. The experimental data of Figure 3 is used to the calculation of the evaluation function Q proposed in this paper, Q changes with k is as shown in Figure 6. Partition algorithms based on double constraint and K-average are used to the experimental data of Figure 3 here.

The simulating result of Figure 6 has shown that the change of Q increases firstly and reduces lately, Q reaches the maximum when $k = 3$ and $k = 5$ in Figure 6(g) and Figure 6(h). It meets the characteristics of the experimental data set in Figure 3. The results illustrate that the optimal partition has biggest difference between each time segment and least difference of one time segment, the calculation of difference between each time segment has weaken its growth in order to get a great value of Q in a reasonable range of k. The result has also proved that Q does not depend on specific partition algorithms.



(e)Figure 3(a) data set (f) Figure 3(b) data set

Figure 5 the tendency of F based on experimental data set



(g) Figure 3(a) data set (h) Figure 3(b) data set

Figure 6 the tendency of Q based on experimental data set

5. Summary

The alarm correlation analysis plays an important role in the fault analysis of the network. Through the time window operation, the original alarms will be tum to the sets of transactions. The ultimate goal of the optimal partition selection is to improve the efficiency of the association rules mining and quickly position even forecast the network faults. This paper presents a new evaluation method of the quality of time segment, it takes advantage of the evaluation function to examine the effectiveness of the algorithm and then finds the optimal partition which improves the preprocessing efficiency of the alarm sequence and provides good data support for subsequent association rule mining.

References

[1]. G.H. Wang, P. Jiang. Survey of Data Mining [J]. Journal of TONGJI University, 2004, 02: 246-252.

[2]. J.F. Li, H.B. Wang. Network Fault Alarm Correlation Analysis Based on Association Rule [J]. Computer Engineering, 2012, 05: 44-46.

[3]. Wang Y, Li G, Xu Y, et al. An algorithm for mining of association rules for the information communication network alarms based on swarm intelligence [J]. Mathematical Problems in Engineering, 2014: 1-14.

[4]. T.Y. Li, X.M. Li. Study of alarm preprocessing based on double constraint sliding time window[J]. Application Research of Computers, 2013,02 : 582-584.

[5]. Y. Yang, F. Jin. KAMEL Mohamed. Survey of clustering validity evaluation [J]. Application Research of Computers. 2008, 06: 1630-1632.

[6]. S.L. Yang, Y.S. Li, X.X. Hu, R.Y. Pan. Optimization Study on *k* Value of *K*-means Algorithm[J], Systems Engineering Theory Practice , 2006,02:97-101.

[7]. Y. Wang, J. Tang, Q.F. Tang, C.Y. Yuan, High efficient K-means algorithm for determining optimal number of clusters [J]. Journal of Computer Applications. 2014,05:1331-1335.

[8]. S.B. Zhou, Z.Y. Yu, X.Q. Tang, New method for determining optimal number of clusters in K-means clustering algorithm [J]. Computer Engineering and Applications. 2010,16:27-31.