Outlier data mining of multivariate time series based on association rule mapping

Yongjun Qin

Department of Mathematics and Computer Technology,
Guilin Normal College,
Guilin, China
Email: teacherqinmail@sina.com

Gihong Min*

Paichai University,
155-40 Baejae-ro, Doma-dong,
Seo-gu, Daejeon, South Korea
Email: ming@pcu.ac.kr
*Corresponding author

Abstract: In the outlier data mining with traditional methods, as the data is complex, the outlier data is not effectively classified, which increase the complexity of data classification and reduce the precision of data mining. In this paper, an outlier data mining method of time series based on association mapping is proposed. By using association rule mapping between datasets, the association rule of datasets is determined. The mining factor and relative error are introduced to improve the precision of data mining. The shuffled frog leaping clustering algorithm is applied to cluster the mining factor. The cluster-based multivariate time series classification is used for classification of clusters based on training set category of time series combined with modified K-nearest neighbour algorithm to achieve classification of time series data and outlier data mining. Experimental results show that running time is only 12.9 s when the number of datasets is 200. Compared with traditional methods, our proposed method can effectively improve the precision of data mining.

Keywords: association rule mapping; multivariate; time series; data mining; k-nearest neighbour algorithm; clustering.

Reference to this paper should be made as follows: Qin, Y. and Min, G. (2020) ‘Outlier data mining of multivariate time series based on association rule mapping’, *Int. J. Internet Manufacturing and Services*, Vol. 7, Nos. 1/2, pp.83–96.

Biographical notes: Yongjun Qin is a Master of Engineering and a Lecturer. He has worked in Department of Mathematics and Computer Technology, Guilin Normal College, since 2007. He mainly engaged in professional teaching and scientific research work in computer application technology. He has a wealth of computer professional teaching experience. His main research directions are data mining with artificial intelligence, data mining algorithms and data mining applications. He has in-depth research and unique insights in the research domains, and published six papers (including *Science and Technology Bulletin* journal), as well as undertook four educational reform and scientific research projects.

Gihong Min is currently a Lecturer in Department of Game Engineering, Paichai University. His research interest is data management and processing, big data processing. He has been engaged in programming management over 10 years, and has rich practical experience and strong scientific research ability. He has published more than ten papers in the provincial-level and above official publications. As an expert, he has been invited to act as reviewers for many respected journals, including *The Journal of Supercomputing*, *Multimedia Tools and Applications*, *Computer and Electrical Engineering*, *IET Image Processing*, and so on. As a project leader, he has been entrusted with many projects entrusted by the government's funds.

1 Introduction

Recently, with the rapid development of computer software and hardware, the ability of people to use information technology to generate and collect data has greatly improved (Moradi and Keyvanpour, 2015; Nguyen and Nguyen, 2015; Liu et al., 2013b). Meanwhile, the development of data acquisition technology, network technology, and computer technology has led to the growth of geometric progression of data and thou-sands of databases have been used in the fields of business management, government offices, scientific research and engineering development (Dang et al., 2016; Ikram and Qamar, 2015). As a result, the data collected in large databases becomes a data grave and develops into the situation of rich data with poor information. How can we not be overwhelmed with these massive and rapidly growing masses of data, how to quickly obtain useful information from massive amounts of data, how to fully increase the utilisation rate of information, and how to solve the contradiction between 'rich data' and 'information shortage'? These are all new topics raised in the 'information age characterised by massive information'. To address the problem, the technology of data mining appears and develops rapidly (Xue et al., 2015; Bian et al., 2016; Liu et al., 2013a). Data mining refers to extraction of the hidden information and knowledge which is not known beforehand but potentially useful, from a large number of incomplete, noisy, fuzzy, and random data. Data mining can help people to extract interested knowledge, rule, and higher level information from the databases. It can be used to analyze the data in different way, so that data can be used more effectively (Palacios and Palacios, 2015; Luna et al., 2015). Association rule mining is used to find the relationship of item sets in a large number of data. Time series is an important data object. It refers to the dataset of records in chronological order, which exists in the fields of aerospace, meteorology, stock market, business, e-commerce, biology, and medicine (Almasi and Abadeh, 2015; Hu et al., 2017). For the problem of poor precision of outlier data mining with the current method, the outlier data mining of multivariate time series based on association rule mapping is researched.

Multi-scale theory is introduced in the field of data mining (Liu et al., 2015). However, the current research on multi-scale data mining is not in-depth. It lacks of universal theory and methods. In order to solve these problems, the multi-scale data mining theory is researched and the scaling-up association rule mining algorithm is proposed. Firstly, based on the theory of layered theory, the definitions of data scale and data scale are given; then, according to the research of multi-scale theory, the essence of multi-scale data mining and the core of research are clarified; finally, the association rule mining algorithm pushed by scale is proposed based on multi-scale data theory. The algorithm processes the frequent item sets in the dataset mining results by using sampling theory and the Jaccard similarity coefficient and realises the transition of knowledge in multi-scale data. Experiments and the analysis are carried out with the synthetic dataset and the demographic dataset from H province. Experimental results show that the algorithm has a high coverage rate, but precision of mining is poor. For the problems of low intelligence and incomplete utilisation of log information in security audit system, a security audit system based on association rule mining is proposed (Xu et al., 2016). The system makes full use of the existing audit logs, combines data mining technology, and establishes a database of user and system behaviour patterns, so that abnormal conditions can be detected in a timely manner and the security of the computer can be improved. Based on the traditional Apriori algorithm, an improved E-Apriori algorithm is proposed. The algorithm can reduce the scope of the set of transactions to be scanned, reduce the time complexity of the algorithm, and improve the operating efficiency. Experimental results show that the recognition ability of attack types for the audit system based on association rule mining is improved by more than 10%. Compared with the classical Apriori algorithm and FP-GROWTH algorithm, the performance of the modified E-Apriori algorithm is improved by 51%, especially in large sparse datasets. However, the precision of data mining is poor. Wang and Zhang (2015) propose an abnormal data mining method based on optimised genetic algorithm for complex network data flow. Firstly, sample complex network data streams and use sample results as sample sets. Then, clustering algorithm is used to obtain cluster centres. All cluster centres form initial populations. Finally, carry out genetic manipulation of the initial population, adaptive adjustment of the number of cluster centres and clusters. This method can better adapt to the high dynamic changes of abnormal data features in complex networks. However, the accuracy of this method is not ideal.

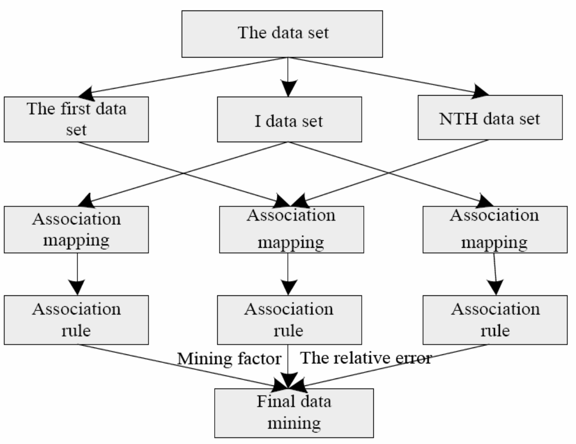
Therefore, based on E-Apriori algorithm, an abnormal data mining method of multivariable time series based on association rule mapping is proposed. This method combines the mapping rules of association rules between datasets and accurately mining abnormal data with multivariable time series.

2 Outlier data mining of multivariate time series based on association rule mapping

2.1 Association rule mapping of dataset

For mining of the topology constructed by data network, the association rule of network dataset is determined with association mapping of network data to improve efficiency of data mining. The mining factor and relative error are introduced to improve the precision of data mining (Wang et al., 2015). The flow chart of data mining with association rule mapping is shown in Figure 1.

Figure 1 Flow chart of data mining with association rule mapping



For the topology of information network, $G(V, E)$is defined as topology, where V is each organisation structure of the network, E is the edge connecting each organisation structure. In $V = (V_1, V_2, \dots, V_n)$, $V = (0 \leq i \leq n)$ denotes the dataset, $V_i = (x_{1i}, x_{2i}, \dots, x_{mi})$, $x_{ij} = (0 \leq j \leq m)$ denotes an effective data in the dataset. Assume association attribute set $(\alpha_{ik}, \beta_{ik}, \theta_{ik})$ denotes the degree of association between the dataset V_i and the dataset V_k, where α_{ik} is size association between datasets, β_{ik} is se-mantic association between datasets, and θ_{ik} is type association between datasets. The association mapping of datasets is defined as follows.

The association mapping of datasets is defined as follows.

Definition 1: Association attribute set $(\alpha_{ik}, \beta_{ik}, \theta_{ik})$ of the dataset V_i and the dataset V_k can represent the degree of association between any data in the dataset.

Definition 2: Association attribute set can be represented by association coefficient matrix. Association coefficient matrix is the average value of association of all data in the two datasets.

$$K_1 = \left\{\begin{matrix} \alpha_{ik} \\ \beta_{ik} \\ \theta_{ik} \end{matrix}\right\} = \left\{\begin{matrix} \alpha_{i1} & \cdots & \alpha_{1k} \\ \vdots & & \vdots \\ \alpha_{1k} & \cdots & \alpha_{i1} \end{matrix}\right\} \left\{\begin{matrix} \alpha_{i1} & \cdots & \alpha_{1k} \\ \vdots & & \vdots \\ \alpha_{1k} & \cdots & \alpha_{i1} \end{matrix}\right\} \left\{\begin{matrix} \theta_{i1} & \cdots & \theta_{1k} \\ \vdots & & \vdots \\ \theta_{1k} & \cdots & \theta_{i1} \end{matrix}\right\} \tag{1}$$

Definition 3: Besides association, there is difference in datasets. Difference coefficient matrix is the reciprocal of association coefficient matrix.

$$K_2 = \left\{\begin{matrix} \dfrac{1}{\alpha_{ik}} \\ \dfrac{1}{\beta_{ik}} \\ \dfrac{1}{\theta_{ik}} \end{matrix}\right\} = \left\{\begin{matrix} \dfrac{1}{\alpha_{i1}} & \cdots & \dfrac{1}{\alpha_{1k}} \\ \vdots & & \vdots \\ \dfrac{1}{\alpha_{1k}} & \cdots & \dfrac{1}{\alpha_{i1}} \end{matrix}\right\} \left\{\begin{matrix} \dfrac{1}{\beta_{i1}} & \cdots & \dfrac{1}{\beta_{1k}} \\ \vdots & & \vdots \\ \dfrac{1}{\beta_{1k}} & \cdots & \dfrac{1}{\beta_{i1}} \end{matrix}\right\} \left\{\begin{matrix} \dfrac{1}{\theta_{i1}} & \cdots & \dfrac{1}{\theta_{1k}} \\ \vdots & & \vdots \\ \dfrac{1}{\theta_{1k}} & \cdots & \dfrac{1}{\theta_{i1}} \end{matrix}\right\} \tag{2}$$

According to association coefficient matrix and difference coefficient matrix, association mapping of the dataset V_i and the dataset V_k is given by equation (3).

$$x_{1i} \to \frac{k_1}{k_2} x_{1k}, \ldots, x_{mi} \to \frac{k_1}{k_2} x_{mk} \tag{3}$$

After obtaining association mapping of the dataset V_i and the dataset V_k, the association rule obtained with cross-correlation matrix is to distinguish the dataset V_i and the dataset V_k. And it is defined by equation (4).

$$f\left(V_i, V_k\right) = \begin{Bmatrix} V_1 & \cdots & V_n \\ \vdots & & \vdots \\ V_n & \cdots & V_1 \end{Bmatrix} \begin{Bmatrix} \alpha_{ik} \\ \beta_{ik} \\ \theta_{ik} \end{Bmatrix} + \begin{Bmatrix} V_1 & \cdots & V_n \\ \vdots & & \vdots \\ V_n & \cdots & V_1 \end{Bmatrix} \begin{Bmatrix} \frac{1}{\alpha_{ik}} \\ \frac{1}{\beta_{ik}} \\ \frac{1}{\theta_{ik}} \end{Bmatrix} \tag{4}$$

And then the two datasets can be classified by association mapping of the two datasets (Luna et al., 2017).

Then the frequency of data mining is obtained by the method of probability estimation. The probability estimation formula is defined by equation (5).

$$P\left(V_i\right) = \frac{\sum_{i=1}^{m} x_i^2}{m \sum_{i=1}^{n} V_i^2} \frac{n}{m(n-1)} f\left(V_i, V_k\right)^{-1} \tag{5}$$

To improve the precision of data mining, the mining factor and relative error are introduced in this paper.

$$P\left(V_i\right) = \lambda^{-1} \xi \frac{\sum_{i=1}^{m} x_i^2}{m \sum_{i=1}^{n} V_i^2} \frac{n}{m(n-1)} f\left(V_i, V_k\right)^{-1} \tag{6}$$

where λ is the mining factor with the value in (0, 1), ξ is the relative error of expected mining probability and the actual mining probability. Only the appropriate value of λ is set can we obtain the maximum probability (Altaf et al., 2017).

2.2 Shuffled frog leaping fuzzy clustering algorithm based on selection and mutation mechanism

To fully reflect the information sharing mechanism of intelligent behaviour and improve the efficiency of local search, the updating strategy of the mining factor is improved.

Inertia weighting ω is introduced to data for coordinating the global and local search ability of the algorithm, it is defined by equation (7).

$$D_i = \omega D_i + rand\left(X_b - X_w\right) \tag{7}$$

where D_i is the moving step of the previous worst frog, and ω is inertia weighting.

Inertia weighting ω represents the trend of frog movement. Near the local optimal solution ω is set to be larger, which is benefit for jumping out of the local optimal solution and avoids premature. Near the global optimal solution, ω is set to be smaller, which is benefit for precisely search near the global optimal solution and finding the global optimal solution. Therefore, in the start of search for flog, ω is set to be larger. As the search proceeds, the algorithm is close to the optimal solution and ω is set to be smaller. According to this, ω is adjusted as linear decreasing by equation (7). $\omega_i(t)$ denotes the inertia weighting of the i^{th} flog for the t times, $\omega_{\min}$ and $\omega_{\max}$ denotes the minimum and maximum weighting, we can obtain equation (8).

$$\omega_i(t) = \left(\omega_{\max} - \omega_{\min}\right)\left[\frac{it \cdot tit - t}{it \cdot tit}\right] + \omega_{\min} \tag{8}$$

where $t = it_j \cdot tit_k$ is the number of iterations in the current subgroup, it_j is the total number of iterations in subgroups, tit_k is the number of the current hybrid iterations, tit is the total number of hybrid iterations, j is a integer of $j \in [1, it]$, k is a integer of $k \in [1, tit]$.

Using the fitness value to measure the pros and cons of the solution, inspired by the roulette strategy in the genetic algorithm, it is proposed to determine the selection probability according to the individual fitness value of the frog. It is calculated by equation (9).

$$P_i = 1 - \frac{f(x_i)}{\sum_{i=1}^{F} f(x_i)} \tag{9}$$

According to equation (9), the smaller the fitness value of a frog individual, the greater the probability of selection; conversely, the smaller the selection probability. In order to play the role of excellent frogs and let the group approach to the optimal solution, a selection mechanism is introduced in the dataset, the selection probability of each frog can be calculated by equation (9), and then we arrange the frogs in descending order according to the selection probability. The probability is that the frog is replaced by the former frog, so that the algorithm proceeds in the direction of the optimal solution (Heraguemi et al., 2016; Yang and Liu, 2014).

If the solution corresponding to the searched location is only a local optimal solution of the optimisation problem, the algorithm appears premature. Therefore, the mutation idea in the genetic algorithm is introduced into the dataset, and the inertia weight ω of the frog and the optimal solution X_g of the frog population are mutated. When the algorithm appears early, let the frog after the mutation operation search in other areas of the solution space to find a better global solution to avoid the algorithm falling into a local optimal solution. The variation equation is defined by equations (10)–(11).

$$\omega' = \omega + \xi_1 N(0,1)\omega \tag{10}$$

$$X_g' = X_g + \xi_2 N(0,1) X_g \tag{11}$$

Here, ω' represents inertia weight value after variation; X'_g represents the optimal solution of the group after variation; ξ_1 and ξ_2 represents variation parameters, $N(0, 1)$ represents random variable with a mean of 0 and a variance of 1. The new algorithm ranks all frogs in descending order of fitness value, uses a greater probability $P_{\max}$ of mutation for the group's former frog, and uses a smaller mutation probability $P_{\min}$ for the group's posterior frogs.

The fitness function measures the individual's degree of adaptation to the clustering problem. For the clustering model, the optimal clustering result corresponds to the smallest value of the objective function, that is, the better the clustering effect, the lower the objective function value, and the greater the individual fitness after clustering (Luna et al., 2016).

Let's define the sample space as $X = \{x_1, x_2, \ldots, x_N\}$, where x_i is d-dimensional vector, A frog in the dataset represents a set of cluster centres and it is defined $C = \{c_1, c_2, \ldots, c_c\}$, where c_j and x_i are coordinated vectors. For the evaluation of each solution (cluster centre) in the shuffled frog leaping algorithm, the individual fitness function is defined by equation (12).

$$f(x_i) = \frac{1}{J+1} = \frac{1}{\sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij} \left\| x_i - v_j \right\| + 1} \tag{12}$$

Here, J is objective function, the smaller J is, the greater $f(x_i)$, and the better the clustering effect is.

The idea of SMSFLA-FCM algorithm is as follows. Firstly, the inertia weighting with linear de-creasing is introduced to the updating strategy of shuffled flog leaping algorithm. The worse flog is replaced by the frog with better fitness selected in a given probability. Each individual of frog has variation with different probability. Secondly, the modified shuffled flog leaping algorithm is to obtain the optimum solution (clustering centre) as the initial clustering centre of FCM algorithm, and then the initial clustering centre is optimised by FCM algorithm. At last, the global optimum solution is obtained. The steps of the algorithm are as follows.

1 Parameter initialisation. Given number of clusters C, permissible error ε, $l = 1$, fuzzy index m, the total number of frog population F, the number of subgroups s, the number of flogs in each subgroup n, the total number of iterations in the subgroup *it*, the total number of hybrid iterations *tit*, variation parameter ξ_1, ξ_2, mutation probability $P_{\min}$, $P_{\max}$.

2 Randomly generate frog population, initialise frog population $v_1, \ldots v_c$, where v_j is a set of arbitrarily generated clustering centre $X = \{x_1, x_2, \ldots, x_N\}$.

3 For each flog, calculate membership matrix U.

4 Calculate the fitness $f(x_i)$ of each flog according to equation (12), and sort frog population in descending order and divide into subgroups according to $f(x_i)$.

5 Dynamically adjust ω, replace the worst flog until to the given total number of iterations *it*.

6 The updated subgroups are mixed to replace the original population. The previous $\frac{1}{4}$ flogs replace the last $\frac{1}{4}$ flogs.

7 According to equations (10)–(11), $P_{\max}$ is used to variation of the previous $\frac{1}{2}$ flogs and $P_{\min}$ to the last $\frac{1}{2}$ flogs.

8 Iterations stop until the largest number of iterations *tit*. The optimal solution is found in the last iteration and the flogs of X_g are outputted to be the set of clustering centre. Otherwise go to 3, $l = l + 1$.

9 Update membership matrix of the flog population.

10 Update the clustering centre and calculate the difference E between two adjacent iterations of membership matrix. If $E < \varepsilon$, stop iterations, otherwise return step 9.

2.3 Cluster-based data classification of multivariate time series

The cluster-based data classification of multivariate time series is proposed on the basis of the above fuzzy clustering algorithm.

With the growth of similar data and the frequent evolution of data characteristics, it is necessary to propose a representative object group which can dynamically describe similar data features, so that it can better express the characteristics of similar data (Pan et al., 2017; Cai et al., 2017).

Definition 4: For the dataset A, if $a_0 = Rep(A)$, a_0 is the representative object of the dataset A, where Rep is the function for solving representative object, which can be the functions of mean, median or mode.

Definition 5: Cluster centre group is a set of several representative objects in the same class of dataset, which is to obtain the minimum distance between the object being represented and the representative object. For the dataset $A = [a_1, a_2, \ldots, a_M]$ in the same cluster, the dataset is divided into K subsets $B = [B_1, B_2, \ldots, B_K]$. Cluster centre group $C = [C_1, C_2, \ldots, C_K]$ is obtained with $c_k = Rep(B_K)$, where $B_i \in A$ and $B_i \cap B_j = \varnothing$.

Cluster centre group is the set of representative objects of data subsets with minimum difference in the same cluster. Compared with the traditional single representative object, it better reflects the data feature and reduce the error of distance of the representative object and the represented object.

$$\sum_{K=1}^{K}\sum_{i=1}^{|B_k|}\left\|b_k - c_k\right\|^2 \leq \sum_{j=1}^{M}\left\|a_j - a_0\right\|^2 \tag{13}$$

where $|B_k|$ is the modulus of the data subset which represents the number of data objects in B_k, b_k is the i^{th} data object in the data subset B_k, that is $B_k = [b_{k1}, b_{k2}, \ldots, b_{k|B_k|}]$.

To better select cluster centre groups, a centre cluster selection method based on neighbouring propagation clustering is proposed. Use the nearest neighbour propagation clustering algorithm to automatically cluster all object sets in the same cluster, generate sub-clusters and obtain the representative object of each sub-cluster. Then, the DBA (dynamic bandwidth allocation) algorithm calculates the mean centre sequence of each sub-cluster using the corresponding representative object as the initial centre sequence. The set of mean centre sequences generated by all sub-clusters is considered as a cluster centre group. The algorithm process is as follows.

Time cluster centre group method based on AP clustering: $C = APCG(A)$

Input: time series datasets in the same cluster $A = [a_1, a_2, \ldots, a_M]$, where a_i represents a kinds of time series;

Output: Cluster centre group $C = [c_1, c_2, \ldots, c_K]$, where c_K represents the k^{th} centre representative object.

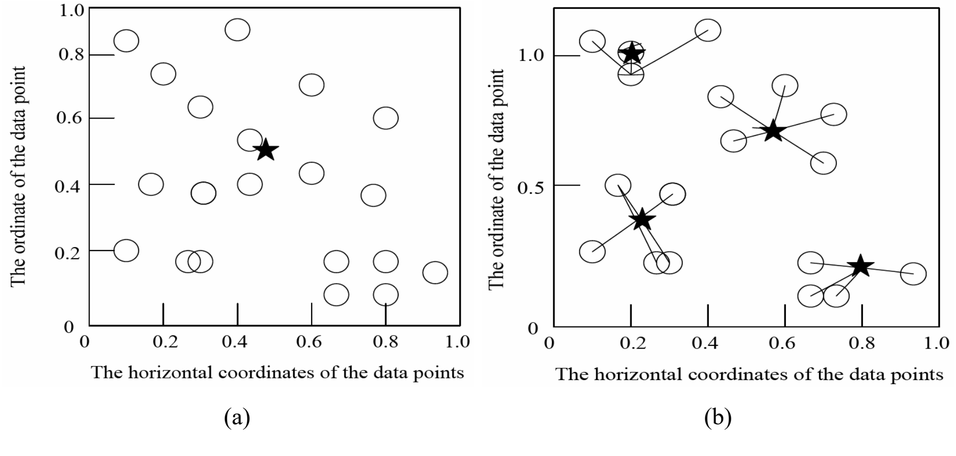
Step 1 The same cluster is divided into k sub-clusters and corresponding representative objects based on AP clustering algorithm. Results can be denoted by $[B, C'] = AP(A)$, where B and C' represents divided sub-cluster sets and representative object sets. $B = \{B_1, B_2, \ldots, B_K\}$ and $C' = \{c'_1, c'_2, ..., c'_K\}$. c'_K represents the k^{th} centre representative object of the k^{th} sub-cluster.

Step 2 Taking c'_K as the initial centre sequence, calculate the central sequence of the corresponding sub-cluster B_k, results can be denoted by $c_K = \text{DBA}(B_k, c'_K)$.

Step 3 Repeat steps 2, calculate the mean centre sequence of all sub-clusters B and obtain the cluster centre group $C = [c_1, c_2, \ldots, c_K]$.

AP clustering can be used to divide adaptively the dataset with the same cluster into several sub-classes, denoted by K classes. For each class, DBA denotes the feature of time series subset.

Figure 2 Representative object based on single mean and cluster centre group, (a) UCO (b) APCG



In Figure 2, square with pentagram represents mean centre, circle with pentagram represents the representative objects generated by AP clustering algorithm, and the start of arrow represents the represented object. If single mean is used to represent all data in the same cluster, as the deviation is large, the representation centre has a weak ability to represent data. Conversely, the object in the cluster centre group represents more similar dataset and the deviation is small. So it has a strong ability to represent data.

The proposed K-nearest neighbour classification method based on cluster centre group uses APCG algorithm to calculate centre group in training set for each cluster. Therefore, a centre group can be used to represent the overall feature for each cluster. The member objects of all centre groups are regard as the new constructed training dataset. For each data object of test set, DKNN is used to achieve classification in the new constructed training set. The algorithm is as follows.

The K-nearest neighbour classification method based on cluster centre group: L = KNN2CG(A, B, K).

Input: training set A, test set B, and the number of nearest neighbours K.

Output: member class label set L in test set B.

1 The training set $A = [a_1, a_2, \ldots, a_N]$ is divided into the corresponding clusters according to the member class label, that is $A = [A_1, A_2, \ldots, A_w]$, where w is the number of classes in A.

2 Each cluster A is calculated to obtain the centre series group by using APCG, that is $C_i = \text{APCG}(A_i)$. Then cluster centre set C is obtained and $C = C_i$.

3 For each data object b_j of the test set, KNN2CG is used to predict the class label in the cluster centre set C, and then $l_j = \text{KNN2CG}(b_j, C, K)$.

4 Repeat the step 3 to obtain predicted class label of data member in all test sets, that is $L = [l_1, l_2, \ldots, l_M]$, where M is the number of members in test set B.

As the new constructed feature training set is far less than the original training set, DKNN can rapidly and effectively classify the time series. For the analysis of time efficiency, the time complexity of KNN2CG method is decided by the learning time T_1 of training set and the prediction time T_2 of test set, and it is defined by equation (14).

$$T = T_1 + T_2 = O\left(N^2\left(t + \frac{m^2}{2}\right) + KMm^2N'\right) \tag{14}$$

As the number N' of members of the new constructed training set is far less than the number N of members of the original training set, the prediction time of the new method will far less than the traditional K-nearest neighbour, that is $KMm^2 N' \prec KMm^2N$. Therefore, from the analysis of time complexity, it can be seen that the new method has better prediction time efficiency.

3 Experimental results and analysis

To verify the proposed method of multi-dimensional data mining based on association rule mapping in bio-information network, the experiment hardware platform is IBM PC with 2.3 GHz CPU, Windows XP operating system, and 4 GB memory and the software is MATLAB 7.0. Random real dataset is used in the experiments.

Figure 3 shows the memory usage for different datasets. The less usage of memory represents the better performance of the data mining algorithm. From Figure 3, it can be seen that the memory usage of the mining algorithm based on association rule mapping is less and the algorithm based on rough set theory and heterogeneous information network is more. Therefore, the data mining performance of the proposed algorithm is better.

Figure 3 Memory usage for different datasets

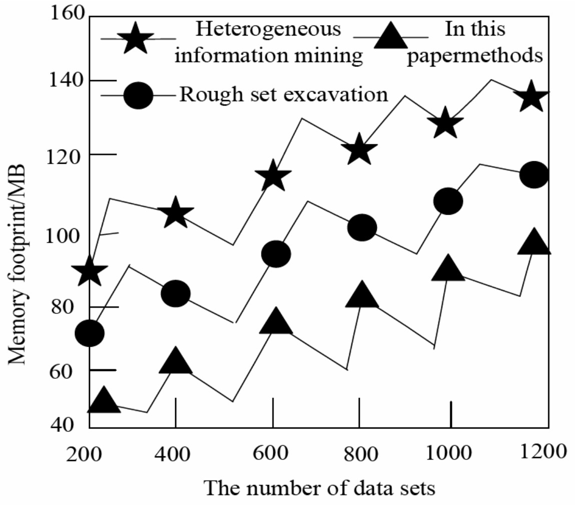
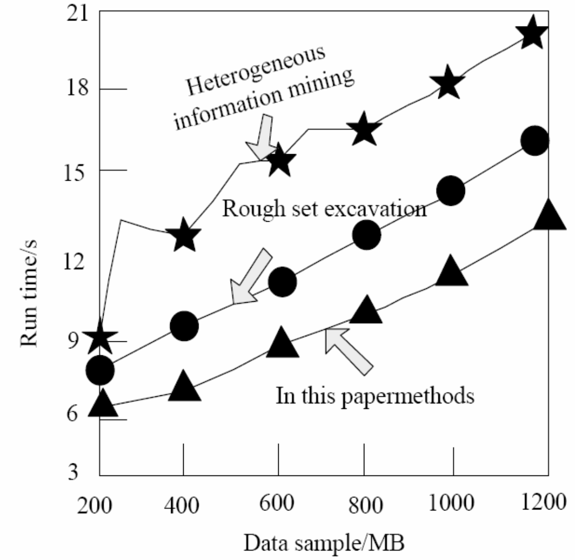


Figure 4 shows the running time of algorithms for different datasets. The running time of algorithms is increased with the increase of the number of datasets. In the case of 1,200 datasets, the running time of the proposed algorithm is 12.9 s, the algorithm based on heterogeneous information is 16.8 s, and the algorithm based on rough set theory is 20.9 s. The less running time represents the better performance of computation of the algorithm, which is suitable for the large scale dataset. Then the precision of outlier data mining is improved.

Figure 4 Running time of algorithms for different number of datasets



For classification of each dataset, the time costs of the two methods are record in the case of different values K of neighbours. The time efficiency of the two methods with different number of neighbours and datasets is shown in Figure 5.

Figure 5 Average consumption of time of the two methods in case of different K and datasets, (a) different number of neighbours (b) different number of datasets

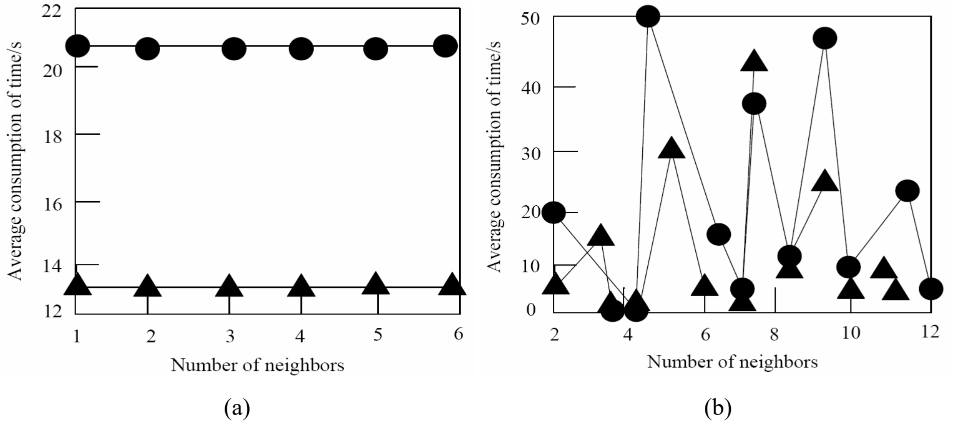


Figure 5(a) shows that the consumption of time of the two methods is increased with the increase of K. It means the time efficiency of the KNN2CG method is better than KNN method in different K values. Figure 5(b) shows that in most of datasets, time efficiency of the proposed method is better than KNN. For the test set, the small test set and long time series data object easily achieve better time efficiency for KNN2CG and effectively improve the precision of outlier data mining.

4 Conclusions

The frequency and precision of data mining can be improved with association mapping between datasets. In this paper, the AP clustering is used to cluster and select the representative object for each cluster in the training dataset. The representative object is regard as initial centre object, and then DBA is used to calculate centre series for each cluster to construct training cluster centre group. Meanwhile, combined with modified K-nearest neighbour method, the classification algorithm based on cluster centre group can achieve better effect and computation performance. The advantages of the proposed method are as follows.

1 Through AP and DBA, the data subset of time series data with very similar shape is described by mean centre series, which reduce the member number of the new training set and improve the performance of computation of the classification algorithm.

2 The centre group provides more detail of overall feature for each cluster. Combined with DTW, mean centre series can better represent the shape feature of the described object and improve the performance of K-nearest neighbour algorithm.

3 K-nearest objects are selected by using average distance. It overcomes the problem that the traditional K-nearest neighbour method is limited to the local optimum. Experimental results show that, compared with the traditional method, the proposed method has better quality of classification and high computation efficiency.

At present, the research on abnormal data mining of multivariate time series is not yet in-depth, and extensive application research is doing. It is not limited to the issues discussed in this paper. In the future, many important theoretical and practical issues need to be resolved. Besides, new research methods need to be explored.

References

Almasi, M. and Abadeh, M.S. (2015) 'Rare-PEARs: a new multi objective evolutionary algorithm to mine rare and non-redundant quantitative association rules', *Knowledge-Based Systems*, Vol. 89, No. 1, pp.366–384.

Altaf, W., Shahbaz, M. and Guergachi, A. (2017) 'Applications of association rule mining in health informatics: a survey', *Artificial Intelligence Review*, Vol. 47, No. 3, pp.313–340.

Bian, Y., Liu, B. and Li, Y. (2016) 'Characterizing network traffic behaviour using granule-based association rule mining', *Networks*, Vol. 26, No. 4, pp.308–329.

Cai, R., Liu, M. and Hu, Y. (2017) 'Identification of adverse drug-drug interactions through causal association rule discovery from spontaneous adverse event reports', *Artificial Intelligence in Medicine*, Vol. 76, No. C, pp.7–15.

Dang, N., Nguyen, L.T.T. and Vo, B. (2016) 'Efficient mining of class association rules with the itemset constraint', *Knowledge-Based Systems*, Vol. 103, No. C, pp.73–88.

Heraguemi, K.E., Kamel, N. and Drias, H. (2016) 'Multi-swarm bat algorithm for association rule mining using multiple cooperative strategies', *Applied Intelligence*, Vol. 45, No. 4, pp.1–13.

Hu, W., Chen, T. and Shah, S.L. (2017) 'Discovering association rules of mode-dependent alarms from alarm and event logs', *IEEE Transactions on Control Systems Technology*, Vol. 26, No. 3, pp.971–983.

Ikram, A. and Qamar, U. (2015) 'Developing an expert system based on association rules and predicate logic for earthquake prediction', *Knowledge-Based Systems*, Vol. 75, No. C, pp.87–103.

Liu, M., Zhao, S. and Chen, M. (2015) 'Scaling-up mining algorithm of multi-scale association rules mining', *Application Research of Computers*, Vol. 32, No. 10, pp.2924–2929.

Liu, S., Fu, W. and Deng, H. (2013a) 'Distributional fractal creating algorithm in parallel environment', *International Journal of Distributed Sensor Networks*, Vol. 9, No. 9, pp.281707.

Liu, S., Fu, W. and Zhao, W. (2013b) 'A novel fusion method by static and moving facial capture', *Mathematical Problems in Engineering*, No. 5, pp.497–504.

Luna, J.M., Cano, A. and Pechenizkiy, M. (2016) 'Speeding-up association rule mining with inverted index compression', *IEEE Transactions on Cybernetics*, Vol. 46, No. 12, pp.3059–3072.

Luna, J.M., Pechenizkiy, M. and Jesus, M.J.D. (2017) 'Mining context-aware association rules using grammar-based genetic programming', *IEEE Transactions on Cybernetics*, doi: 10.1109/TCYB.2017.2750919.

Luna, J.M., Romero, C. and Romero, J.R. (2015) 'An evolutionary algorithm for the discovery of rare class association rules in learning management systems', *Applied Intelligence*, Vol. 42, No. 3, pp.501–513.

Moradi, M. and Keyvanpour, M.R. (2015) 'An analytical review of XML association rules mining', *Artificial Intelligence Review*, Vol. 43, No. 2, pp.277–300.

Nguyen, L.T.T. and Nguyen, N.T. (2015) ‘Updating mined class association rules for record insertion’, *Applied Intelligence*, Vol. 42, No. 4, pp.707–721.

Palacios, A.M. and Palacios, J.L. (2015) ‘Genetic learning of the membership functions for mining fuzzy association rules from low quality data’, *Information Sciences*, Vol. 295, No. C, pp.358–378.

Pan, Z., Liu, S. and Fu, W. (2017) ‘A review of visual moving target tracking’, *Multimedia Tools and Applications*, Vol. 76, No. 16, pp.16989–17018.

Wang, H. and Zhang, C-Y. (2015) ‘Differences between network data mining algorithm based on improved genetic algorithm’, *Computer Simulation*, Vol. 32, No. 5, pp.311–314.

Xu, K., Gong, X. and Cheng, M. (2016) ‘Audit log association rule mining based on improved Apriori algorithm’, *Journal of Computer Applications*, Vol. 36, No. 7, pp.1847–1851.

Xue, C., Song, W. and Qin, L. (2015) ‘A mutual-information-based mining method for marine abnormal association rules’, *Computers and Geosciences*, Vol. 76, pp.121–129 [online] https://doi.org/10.1016/j.cageo.2014.12.001.

Yang, G. and Liu, S. (2014) ‘Distributed cooperative algorithm for k-M set with negative integer k by fractal symmetrical property’, *International Journal of Distributed Sensor Networks*, Vol. 10, No. 5, p.398583.