Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (11): 1980-1983, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.1980.1983

Urban Tunnel Clean up the Dirty Data Based on Clara Optimization Clustering

[1,2]KunHao Tang, [1]Luo Zhong, [1]Lin Li and [1]Guang Yang
[1]Department of Computer Science and Technology, Wuhan University of Technology, 430070, China
[2]Department of Computer and Information Science, Hunan Institute of Technology School, 421002, China

Abstract: Along with the construction of intelligent city accelerates, the popularity of wireless intelligent device, the data we can collected increase exponentially. In terms of Intelligent tunnel, this study collects the noise pollution of real-time data, by comparing the advantages and disadvantages between the traditional clustering k-means algorithm and traditional Clara clustering algorithm and proposes a cleaning for improved data optimization algorithm based on randperm function, called RICA, in order to improve accuracy and reduce the running time as the first target. Random functions of the algorithm proposed is effective for tunnel internal moment change circulation of large amount of data to make corresponding changes. The experimental results show that among the same set of data, the algorithm has shortened the time ,increased the cyber average phase and accuracy effectively, so that it is a reliable and improved method for data processing.

Key words: Clustering, the average dissimilarity degree, the improved Clara algorithm, k-means algorithm

INTRODUCTION

Along with the construction of intelligent city accelerates, the rapid development of computer technology and the popularity of the Internet of things, a variety of wired and wireless device data appear and change every time. And our life is filled with more and more data. How to process data repetitiously and analysis thoroughly from the data ocean, for enabling us to gain knowledge effectively and dig out the key information needed to successfully becomes a research focus at the current time in data mining (Zhong *et al*., 2012). In the whole process of data mining, data preprocessing time accounted for about 60% and the final data mining work only takes up about ten percent of the time. The number of data preprocessing involves is so big that it must be achieved through the program and algorithm (Wan *et al*., 2012).

The purpose of data preprocessing is using of relevant technology to extract.

Funding: Wuhan technology department, innovation project (2013070204005); Hengyang technology department (2012KG76); Hunan education department (11C0366); Hunan technology department (2013GK3037) data from various data sources in order to convert the dirty data or clean up it (Zhang, *et al*., 2012). Commonly used technical means include Binning smooth, regression and clustering.

In general, data clustering technology can divided the collection of data objects into several clusters consisted of similar data while cleaning up abnormal data and dividing the normal data. As one of the important technology of data mining, data clustering technology has obtained remarkable achievement in the area such as market analysis, financial investment, health care and other areas of the practical application.

THE OVERALL FRAMEWORK OF TUNNEL MONITORING DATA

The city tunnel monitoring data include a variety of forms. Because the tunnel monitoring system demands real-time strictly, the collected data storage density is larger and the data storage capacity is more. So that a large amount of data can only provide effective data mining after data cleaning (Thakran and Toshniwal, 2012).

Table 1 is the city tunnel monitoring data preprocessing frame. This study take the analog data measured by these city tunnel as the research object, absorb the experience of the traditional clustering mode and combine respective advantages for processing large data quantity. It build a new algorithm based on Clara optimization clustering.

Corresponding Author: KunHao Tang, Department of Computer Science and Technology,Wuhan University of Technology, 430070, China Tel: 13975466676

Table 1: The sources and pre-processing of city tunnel data

Data sources	Data types	Pre-processing method	Anticipated goal
Tunnel car inspection data	Time Traffic Driving speed Driveway occupancy rate	Clustering and SOM	Delete illogic dirty data and analysis the reason is accident or equipment damage

Clara algorithm optimization: Clustering is a statistical analysis method for the study of classification problems. It Clustering is based on similarity, by which similarity can be achieved among data within the cluster and difference can be seen among clusters. Clustering can be divided based on the division, density, probabilistic models and other types. This study studies the clustering method based on the division (Ming Xiu Duan, 2010).

Description of clara algorithm: Clara algorithm is a sampling-based approach, it is able to handle data in large amount. In order to achieve better clustering effect, Clara algorithm should take multiple samples from actual data and replace the corresponding center with the PAM on each sample, then output the best clustering case. Specific steps are as follows:

Step 1: Randomly draw a 40+2 k sample of objects from the entire data set and use the PAM replacement strategy to find the center of the k samples

Step 2: As to the entire data set for each object Q which k center is most similar to Q should be found, then assign Q to it and record its corresponding class

Step 3: Calculate the average dissimilarity obtained in cluster (Ben-Arieh and Gullipalli, 2012). If the result turns out to be less than the current minimum, then convert it to be the new current minimum and retain the k center obtained in the step 2 as the best center collection so far

Step 4: Repeat step 2 to Step 4 and output the best result after several tests

Description of RICA (randperm invoking clara algorithm): RICA is the abbreviation of Randperm Invoking Clara Algorithm (Clara algorithm based on random and distinct function). Because the Clara random sampling exist the problem of repeated data and circulation while the existence of circulation in the MATLAB will seriously affect the execution efficiency, we should take steps to improve. Our experiment use the randperm function to get the distinct sample so that we avoid the cycle test and improve the efficiency of Clara algorithm. RICA specific steps are as follows:

Step 1: Wipe off the non-numeric and missing data in the whole data sets

Step 2: Take the whole data sets as the object, get 40+2×k random non-repetitive data as samples by randperm function. And find k centers in the samples by using PAM replacement strategy

Step 3: For the arbitrary object Q in the whole data sets, judge the center which is nearest the object Q. And divide the Q into the Cluster that the nearest center is in. Take down the number of cluster

Step 4: Calculate the average dissimilarity of the case in step 3 and compare it to the current minimum. If the dissimilarity is smaller, convert it to the current minimum and keep in the centers as the best cluster centers until the better case happen

Step 5: Repeat step 2 to step 4 to perform several experiment and get the best result as the output.

Data cleaning based on the clustering: After using RICA algorithm for clustering, we can not only get the k clusters, but also can obtain the relationship between each point in the same cluster, including the dissimilarity (Euclidean distance) between the cluster center and every point in the cluster. Except, we can also get the dissimilarity in different dimension so that we can easily wash out the dirty data. Specific steps are as follows:

Step 1: Set the maximum and minimum values for all the dimension of record,that is,set the limits

Step 2: Find out the maximum value of the distance between the various data in each dimension for each cluster center:

$$d_{i,j} = \max\{|a_{i,j}-m_j|, |a_{i,j}-n_j|\} \quad (1)$$

The 'i' is the number of cluster, i = 1, 2,..., k; The 'j' is the number of dimension, j = 1, 2,..., l, the 'l' is the limit of dimension. $a_{i,j}$ means the value of dimension j in the cluster i.m_j and n_j stand for the maximum and minimum value of data in dimension j

Step 3: According to the most value of the distance between each dimension data and edge,and the weight of each dimension to estimate reasonable range of the dissimilarity between the normal point and the cluster center:

$$\text{Distance}(i) = \sqrt{\sum_{j=1}^{l}(d_{i,j}^2 \times Q_j)} \quad (2)$$

The 'i' is the number of cluster, the 'j' is the number of dimension. And Qj stand for the weight of dimension j.$Q_j = t_j/l$, t_j is the appear times in the dimension j

Step 4: According to the reasonable range distance (i), find out the dirty data(the corresponding dissimilarity is beyond the range)

Step 5: Test the dirty data and judge the fault type according to the number of dimension which is beyond the limit. (If the number is one, it means equipment damage, or it stands for accident)

EXPERIMENTS

The experimental evaluation

The average dissimilarity: The calculation formula of Euclidean distance is as follows:

$$d(x, y) = \sqrt{\sum_{i=1}^{m} (x_i, y_i)^2} \quad (3)$$

the m is dimension of data.

Silhouette: Silhouette is a kind of used for evaluation of clustering validity evaluation index. Silhouette value s(i) is defined as:

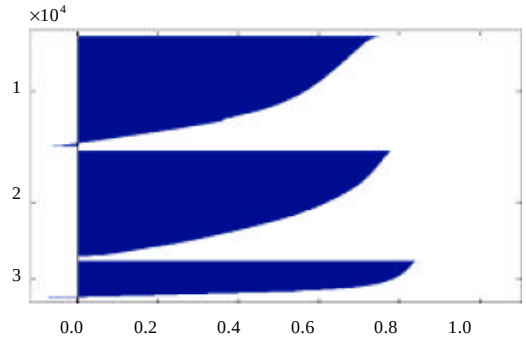

Fig. 1: k-means outline figure

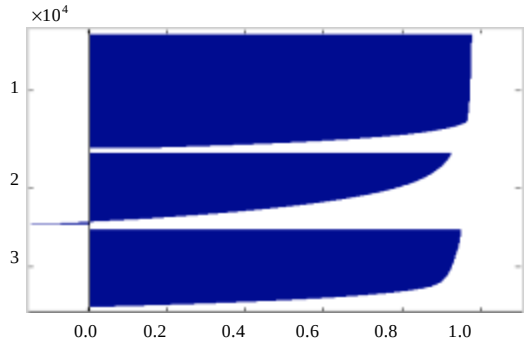

Fig. 2: Clara outline figure

$$\frac{b(i) - a(i)}{\max\{b(i), a(i)\}} \quad (4)$$

a(i) is the average distance between the sample i and other sample in the same cluster. And b(i) is the minimum value of the distance between sample i and all the samples included in the Cluster Xt (t = 1, 2,..., k; t ≠ j).

The experimental results and analysis: The study uses the method of k-means and Clara Algorithm to cluster and analysis the datasets and compare the results. The experimental is as follow.

Silhouette: Each point represents a record, plotted by Silhouette on the horizontal axis and cluster on the vertical. The silhouette index of Clara algorithm is better than k-means a lot and already very close to 1.The case that Silhouette index is close to -1 is race .It means that it is rare to divide data into wrong cluster.

Dissimilarity and run time: The purpose of clustering is to make the data in same cluster is similar to each other and the date in different cluster is dissimilar to each other. Using the average Euclidean distance as the indicator to measure the average dissimilarity can show the level of similarity clearly. The smaller the average dissimilarity is, the more ideal the experimental result will be. Moreover, it only costs less than 20 sec to deal with the 42993 records using the k-means or Clara algorithms (Wang and Zhong, 2009).

The comparison between the two algorithms is as follows.

Cluster results: In order to see the clustering results clearly, we turn the multidimensional data into the two-dimensional. We can find the k-means and Clara algorithm can have very good clustering effect by mapping corresponding two-dimensional scatter plot figure. The figures are as follows:

Dirty data cleaning: We get three reasonable clusters by using RICA algorithm with 42993 tunnel data

Table 2: The average dissimilarity of algorithms

Algorithm	Cluster	Dissimilarity	Dissimilarity	Run time
k-means	3	9.9983e+005	23.2556	8.5460 sec
k-means	4	9.7394e+005	22.6535	10.7040 sec
k-means	5	9.9777e+005	23.2077	14.9530 sec
Clara	3	9.9528e+005	23.1498	7.9700 sec
Clara	4	9.7098e+005	22.5846	10.0530
Clara	5	9.4098e+005	21.8868	13.8270

Table 3: Deal with dirty data

No.	Clustering No.	Dissimilarity	Reasonable range	Wrong dimension	Reason
2	1	132.0446	115.1515	One	Equipment damage
3	1	231.0190	115.1515	Two or more	Accident
4	1	329.6152	115.1515	Two or more	Accident
11475	2	108.0017	101.5867	One	Equipment damage
38956	3	305.4176	102.6712	Two or more	Accident
38976	2	305.8578	101.5867	Two or more	Accident
41631	2	113.4741	101.5867	One	Equipment damage
41651	1	122.8652	115.1515	One	Equipment damage

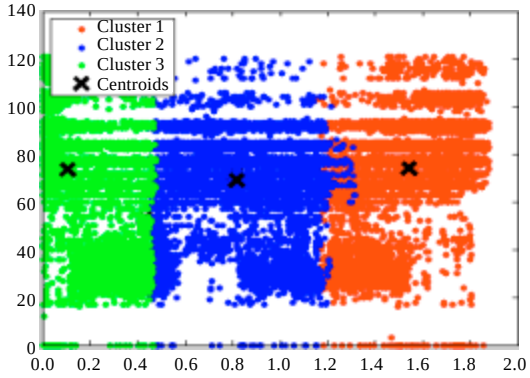

Fig. 3: k-means scatter plot figure

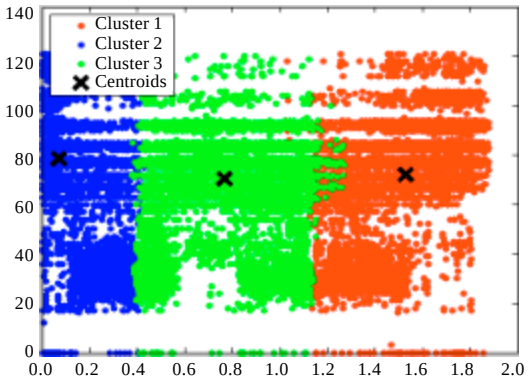

Fig. 4: Clara scatter plot figur

records.After dealing with the clustering results,we can find the dirty data and clean it.The experimental results are as follows:

CONCLUSION

The experimental results show that the RICA clustering algorithm is very effective for the large-amount data, it can be effective to clean our data.It not only gathers the similar data together and digs out the potential relationship among data, but also finds abnormal data quickly and efficiently.So that it is convenient for us to analysis of the accidents in tunnels.

REFERENCES

Ben-Arieh, D. and D.K. Gullipalli, 2012. Data envelopment analysis of clinics with sparse data: Fuzzy clustering approach. Comput. Ind. Eng., 63: 13-21.

Duan, M.X., 2010. Improved CLARA clustering algorithm combined with SOFM. Comput. Eng. Applic., 22: 210-212.

Thakran, Y. and D. Toshniwal, 2012. Intelligent Systems Design and Applications (ISDA). Intelligent Systems Design and Applications (ISDA). Proceedings of the 12th International Conference on Digital Object Identifier, November 25-27, 2012, pp: 947 952.

Wan, M., C. Wanga, L. Lia and Y. Yanga, 2012. Chaotic ant swarm approach for data clustering. Applied Soft Comput., 12: 2387-2387.

Wang, Q.B. and L. Zhong, 2009. Urban highway tunnel monitoring system integration method. J. Wuhan Univ. Technol., 31: 729-732.

Zhang, Q.J., L. Zhong, J.L. Yuan and Q. Wang, 2012. Design and realization of a city tunnel environment monitoring system based on observer pattern. Proceedings of the 2nd International Conference on Electric Technology and Civil Engineering, January 2012, Washington, DC, USA., pp: 808-811.

Zhong, L., L. Li, J.L. Yuan, H.X. Xia, Q.B. Wang and Q.J. Zhang, 2012. Urban tunnel monitoring data storage method based on XML and its system. China, Invention patent, 201110458995.2.