Scholars Journal of Engineering and Technology (SJET) **ISSN 2321-435X (Online)**
Sch. J. Eng. Tech., 2015; 3(6):595-599 **ISSN 2347-9523 (Print)**

(An International Publisher for Academic and Scientific Resources)
www.saspublisher.com

Research Article

An outlier detection algorithm based on clustering

Hongbo Zhou[1], Bingbing xu[2], Juntao Gao[3]
[1]Northeast Petroleum University, Institute of Computer and Information Technology, Daqing, Heilongjiang, China
[2] Information Center, Well Testing and Perforating Services of Daqing Oilfield Company Limited, Daqing, China
[3]Northeast Petroleum University, Institute of Computer and Information Technology, Daqing, Heilongjiang, China

***Corresponding author**
Hongbo Zhou
Email: jiessie9@126.com

Abstract: Outlier detection is a very important type of data mining, which is extensively used in application areas. The traditional cell-based outlier detection algorithm not only takes a large amount of time in processing massive data, but also uses lots of machine resources, which results in the imbalance of the machine load. This paper presents an distance-based outlier detection algorithm. These experiments show that this improved algorithm is able to effectively improve the efficiency of the outlier detection as well as the accuracy.
Keywords: outlier detection; Euclidean distance; cluster center; maximum distance principle; minimum distance principle

INTRODUCTION

Outlier detection is a significant problem in the field of data mining [1]. Due to its importance, several outlier detection approaches have been proposed for data mining. These include the nearest-neighbour based [2], density based[2], clustering based[2] and distance based[2]. Due to the increasing usage of automatic data collection devices, i.e., sensors, RFIDs, etc., measurements contain inherent uncertainty. Therefore, outlier detection from uncertain data is gaining popularity and a lot of researchers are focusing on it.

We focus on distance-based outlier detection algorithm because distance-based approaches are the simplest and the most commonly used. On the one hand it is proved to be an effective method to find outliers, and can be applied to various forms of data. In the actual dissertation we started by presenting a theoretical overview of outlier detection methods and data mining techniques. In this paper, we have proposed a new efficient outlier removal method.

RELATED CONCEPTS

Euclidean distance [3]

Typically spatial clustering algorithm is built on the basis of various distances, such as the Euclidean distance, Manhattan distance [3] and Ming Cowes distance [3]. Among them, the most commonly used is the Euclidean distance, we use Euclidean distance as a standard clustering to analyze the improved k-Means algorithm.

There are two points $i=(x_{i1},x_{i2},\ldots,x_{in})$ and $j=(x_{j1},x_{j2},\ldots,x_{jn})$,which are two n-dimensional data object, then the Euclidean distance between them is defined as follows:

$$d_{ij} = \sqrt{(X_{i1} - X_{j1})^2 + (X_{i2} - X_{j2})^2 + \ldots + (X_{in} - X_{jn})^2} \quad (1)$$

Cluster center [4]

The average of cluster data objects is called clustering center of the cluster, the nearer the cluster center between the two clusters, the more similar the two clusters, the farther the cluster center, the less these two clusters similar. Cluster center of one cluster is calculated as follows: Let n data object containing the data set $X=\{x_1,x_2,x_3,...,x_n\}$, respectively, the cluster center $z_1,z_2,\ldots,z_k$.

$$Z_i = \frac{1}{C_i} \sum_{j \in n_i} j \quad i = 1,2,\ldots,k \quad (2)$$

Maximum distance principle

It is a method of selecting the cluster centers. In the division process of clustering, the cluster center as a reference point to a cluster represents a local division of clustering. A good clustering division between the cluster having the largest difference, and the division in the cluster having the maximum similarity. In the K-means algorithm, the similarities and dissimilarities between the two clusters is evaluated by the value of Euclidean distance. Euclidean distance is smaller, the greater the similarity, dissimilarity smaller. Conversely, the greater the Euclidean distance, the smaller the similarity, the greater dissimilarity. In order to ensure maximum similarity in same clusters and maximum dissimilarity between different clusters, different cluster centers should be possible to maximize the distance.

Minimum distance principle

It is a cluster partitioning method of clustering points. In the division of the clustering process, as a reference point to the cluster center of the cluster represents a local division of clustering, if a clustering is divided into multiple clusters, a distance between any cluster point to be divided and clustering cluster center is minimum, the clustering point will be divided into the cluster including its cluster centers.

For example: a cluster with n data points, $X=\{x_1,x_2,...,\ x_n\}$,$k\in[1,n]$,is divided into k clusters, cluster centers for the i-th cluster x_i,$i\in[1,n]$.In the process of division of clustering, Euclidean distance is compared in turn between data points in X and of cluster centers of k clusters, the point is divided into the cluster which nearest cluster center from the point.

ALGORITHM PRINCIPLE

The principle of clustering algorithm is to make the minimum similarity within the data object and maximum similarity between the data object. In the K-means algorithm, usually consists of a cluster center represents a class,to determine the optimal number of clusters by the number of cluster centers. Therefore, to get a better clustering results,you should make the distance between each cluster center as large as possible. In view of this, the choice of the initial cluster centers, focusing on the distance between them is as large as possible. Taking into account the relatively large data sets are often used method of dividing one by one to find the cluster center, the first to be clustering in the data set to identify two data objects farthest from the initial cluster centers as two to two poly class centers as the basis, the remaining data object divided by the minimum distance from the class to its nearest cluster center belongs to them, to be divided is completed, and then find out the farthest two data objects are in this category, taking the maximum that the distance to these two data objects such as clustering center divide, and so constantly draw farthest category two data objects, until clustering is convergence.Due to constantly divide the data set and data farthest object as initial cluster centers, and therefore, there is a greater distance between any two cluster centers obtained by this method.

Our algorithm in based on the idea of filtering the data after executing the clustering process. The proposed method has four main processing stages.The first step in accordance with the principle of the maximum distance to find accumulation point, as the focal point of the cluster center, the second step according to the minimum distance principle, to choose the cluster center as the center point of the cluster is divided into several sub-clusters, the third step, check each sub-poly number of cluster points, if a sub-cluster number of points in 1-2 poly(you can manually set) between considers the cluster center is an outlier. The fourth step, the focal point of the selected outlier removed and isolated points to add to a collection of isolated points to go. Back to the first step execution, until the number of isolated points does not increase so far.

ALGORITHM PROCESSES

Clustering data sets $N=\{x_1,x_2,x_3,x_4,\ldots,x_{n-2},x_{n-1},x_n\}$ has n data objects, each data object has attributes t, i-th data object can be expressed as $x_i=\{x_{i1},x_{i2},x_{i3},x_4,\ldots,x_{it}\}$, algorithm is as follows：

Step 1、Initialization data set Y,G and W, the number of clusters k, wherein Y is stored in the divided cluster centers, G is stored Outlier point, W is stored no clustering data points, k is the number of clustering and r is the number of cluster data in G.$Y=\{\Phi\}$,$G=\{\Phi\}$,$W=\{x_1,x_2,x_3,x_4,\ldots,x_{n-2},x_{n-1},x_n\}$, set k=1, r=0.

Step 2、Calculating Euclidean distance between any two cluster points in W, and choosing two cluster points p_1 and p_2 in W which has a maximum Euclidean distance.

Step 3、Adding p_1 and p_2 to Y, while p_1 and p_2 is removed from W.$Y=Y\cup\{p_1, p_2\}$.set k=k+1,$W=\{X_i \mid X_i \notin Y\}$,$i\in[1,n]$.

Step 4、Regarding every point in Y as cluster point, these cluster points in W will be divided into the clustering which is located closer to the cluster centers according to the principle of minimum distance, until the division is completed.

Step 5、Each cluster point in Y is analysis number of cluster points divided into sub clusters. If number of cluster points around cluster point in Y is too small, as only 1-2(the number can be artificially defined),the cluster point is considered a outlier point, the point is added to G while the point is removed from Y and W.set r=r+1.For example, if p is a cluster point in

Y,number of cluster points around p is too small, as only 1-2, the cluster point of p is considered a outlier point. set r=r+1,G=G∪{p},Y=Y-{p},W=W-{p}.

Step 6、Set Y={Φ} and k=1,cycle turn step 2,and compare the value of r around two times calculated, if r corresponding around two times is different, cycle turn step 2,otherwise cycle ends.

The cluster points in G is outlier point, r is the number of outlier points, outlier points are removed to G from W,Y=Y-G.W={$X_i|X_i \notin G$},i∈[1,n].The value of k is the number of clusters, and the effective number of clustering points is n-r,set n=n-r.

SIMULATION ANALYSIS

We manually configured by a set of data to test the effectiveness of the algorithm, table 1 is spatial coordinates of study Object.Figure 1 is a study of the distribution of spatial object region, a total of eleven spatial data points, the spatial coordinates as shown in Table 1.Now using the k value of spatial clustering optimization algorithm presented above, taking into account the experience of the rule kmax≤$\sqrt{n}$ [5,6],there should be: kmax≤$\sqrt{11}$ = 3.3166, kmax can only take integer,and therefore the range of k can be reduced to k1=1, k2=2, k3=3, solving steps are as follows:

Table-1: Spatial coordinates of study object

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11
x	1	2	3	3	4	2	8	9	10	2	7
y	1	2	1	4	5	5	3	2	3	9	10

In 11 cluster points, maximum distance is 10.8167, which exist between p1 and p11.Y={p1,p11}. Regarding every point in Y as cluster point, these cluster points in W will be divided into the clustering which is located closer to the cluster centers according to the principle of minimum distance, until the division is completed. The sub cluster is {P1, P2, P3, P4, P5, P6, P8},which cluster center is p1,another sub cluster is {P11, P7, P9, P10},which cluster center is p11.

Because number of cluster points of the two sub clusters divided is larger than 2, there is no outlier point in the division.Euclidean distance is calculated between any two cluster points in the two sub clusters, and choose two cluster points P1,P8 and P9,P10,which has a maximum Euclidean distance.Because the distance between P1 and P8 is smaller than the distance between P9 and P10,P11 is removed from Y and add {P9,P10},Y={ P1,P9,P10}.Regarding every point in Y as cluster point, these cluster points in W will be divided into the clustering which is located closer to the cluster centers according to the principle of minimum distance, until the division is completed. The division is Fig 3.

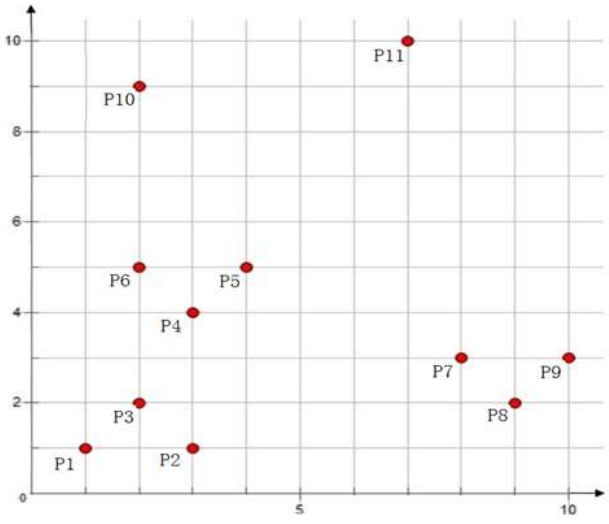

Fig 1 spatial distribution of data points

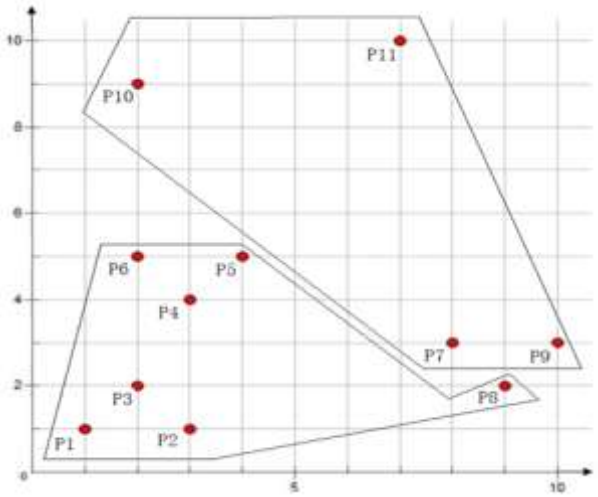

Fig-2: data points are decomposed into 2 sub clusters

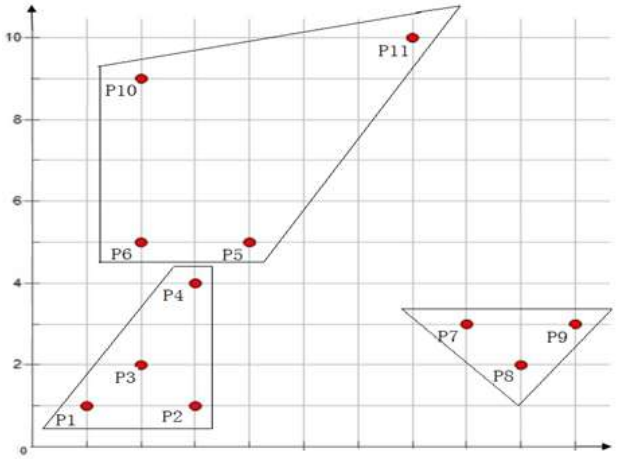

Fig-3: data points are decomposed into 3 sub

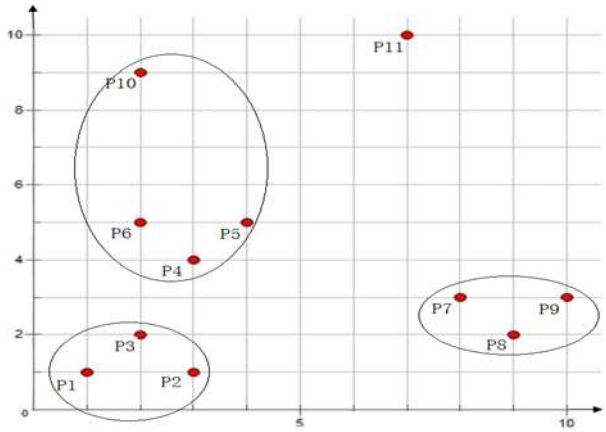

Fig-4: the first isolated point

In the three sub cluster, maximum distance is 7.071 between P6 and P11, P10 is removed from Y, and P6 and P11 is added to Y.Y={P1,P9,P6,P11}. Regarding every point in Y as cluster point, these cluster points in W will be divided into the clustering which is located closer to the cluster centers according to the principle of minimum distance, until the division is completed. The division is Fig 4.

Each cluster point in Y is analysis number of cluster points divided into sub clusters. Because If number of cluster points around P11 is zero, P11 is considered a outlier point, set r=r+1,G=G∪{p},Y=Y-

{p},W=W-{p}.Cycle turn step 2,until r is not different around two times calculated. At the end, P10 can be obtained as an outlier point.

CONCLUSION

Outlier detection is an extremely important problem with direct application in a wide variety of domains. Even though the problem of outlier detection has been studied intensively in the past a few years, outlier detection problem is still a not well formulated problem. In this dissertation we have discussed the different ways in which the problem has been formulated in literature. Every unique problem formulation entails a different approach, resulting in a huge literature on outlier detection techniques. Some researches saw as statistical problem while other tried to find some distance metrics to solve it while other exploited clustering techniques.

This thesis proposes an outlier detection method that provides efficient outlier detection and data clustering capabilities that are noise effective. Our method takes advantage of the data clustering process to filter outliers which enables us to solve two problems in the same time in a reasonable.

ACKNOWLEDGMENT

The paper is supported by the Education Department of Heilongjiang province science and technology research projects (No.1253G014).

REFERENCES

1. Mac Queen J; Some methods for classification and analysis of multivariate observations[C]. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967; 281-297.
2. Jain A K, Dubes RC; Algorithms for Clustering Data. Englewood Cliffs, NJ, USA: Prentice-Hall, 1988.
3. Xu R, Wunsch D; Survey of clustering algorithm. IEEE Trans. Neural Networks, 2005, 16(3): 645-678.
4. Milligan G W, Cooper M C. An examination of procedures for determining the number of clusters in a data set. Psy-chometrika, 1985, 50: 159-179.
5. UCI Repository of machine learning databases and domain theories[EB/OL].FTP address: ftp://ftp.ics.uc.i edu/pub /machine-learning-databases
6. Xie X L, Beni G. A validity measure for fuzzy clustering. IEEE Trans. Pattern Anal. Machin. Intell, 1991, 13(8): 841-847.