Data Mining Applied to Oil Well Using K-means and DBSCAN

Chang Lu , Yueting Shi, Yueyang Chen*, Shiqi Bao, Lixing Tang
School of information and electronics
Beijing Institute of Technology
Beijing, China

Abstract—**Oil is essential to our life mainly in transportation, and thus the productivity of oil well is very important. Classification of oil wells can make it easier to manage wells to ensure good oil productivity. Machine learning is an emerging technology of analyzing data in which cluster is a good way to do classification. The paper will apply two kinds of cluster method to the data from Dagang oil well and then do analysis on not only the classification results but also the choice of method for future analysis.**

Key words--K-means; DBSCAN; cluster; PCA; oil well

I. INTRODUCTION

Machine learning and big data is an emerging synthetic analysis method which can help process large amount of data and do analysis on them through programming. In this paper, we firstly introduce two classification methods and their basic theories. Then how we do data pretreatment and adjustment is presented. At last, we do analysis on the data, show the results using different algorithms, compare the results and think about the differences to explore which classification method should be used in the specific occasion.

II. THEORY

Cluster [1] is a kind of unsupervised learning algorithm. That is to say, the samples do not have an implicit label but only the features which can tell the differences between the samples. The aim of clustering is to find the differences between the points by giving labels to every point so that only the similar points can be divided into the same group which have the same label. The principle is just like the example: the stars in the universe can be presented as a point set (x,y,z) in the 3-dimension space. The distance between the stars in the same cluster is small while that of different star clusters is far.

We will talk about the two clustering methods firstly.

A. K-means clustering

K-means clustering[2] is a method of vector quantization, aiming to partition n data points into k clusters in which each point belongs to the cluster with the shortest distance to mean value, serving as a basic rule of the clustering.

The specific algorithm is shown as below:

- Randomly choose k points to be initial centroids;
- Repeat the following steps until convergence or reach the max iteration times {

For every sample i, calculate the cluster it belongs to

$$c^{(i)} := \arg\min_j \left\| x^{(i)} - \mu_j \right\|^2$$

For every cluster j, recalculate the centroid of the cluster:

$$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}} \}$$

For the equations above, $c^{(i)}$ represents the label of cluster whose centroid is closest to the No. i sample, and the value is a number in the range of one to k. The mean value represents the guess for centroid of each cluster.

It is also important to in advance learn the choice of parameters [3-6] to get good results:

- The number of cluster: one way is to make use of the hierarchy clustering method to firstly test the data and then set the cluster number of K-means to be the same. Also, we can do resampling on the data to get two subsets, and then do similar clustering on the two subsets of data to get the result with k clusters. The similarity of the point cluster distribution is important to decide whether the k value can be used or not and we should also try many times to find the best result.
- The place of centroid: we can choose centroid randomly and do the process several times. At last, choose the best initial condition. An alternative way is we may again use hierarchy method to calculate firstly and use the calculated result. Maybe cross validation is also a good way to analyze.

B.DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm proposed by Martin Ester etc. in 1996.[7] [8]It is a density-based clustering algorithm: given a set of points, it gathers points together in a way that closely distributed points packed together (points with many nearby neighbors) and have the same label.

978-1-5090-3555-7/16 $31.00 © 2016 IEEE
DOI 10.1109/CCBD.2016.43

Points marked as outliers lie away from the existed clusters (whose nearest neighbors are too far away). DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature.

It is also important to note that for DBSCAN, there is no need to set the number of clusters for the algorithm does the classification only according to distribution of the points. Because of the characteristic above, we need firstly to set two parameters to tell how the points distribute can be called the same cluster: Eps(the maximum radius of each cluster) and MinPts(the smallest number of points in the same cluster and the number must be larger or equal to one). We could retrieve all points that are density-reachable based on the set parameters.

The process is shown as fellow: firstly, we should detect the point p which has not been divided into any group, and calculate the number of points around it, which means that the points' distance to it is smaller than Eps). If the number is larger than MinPts, a new cluster will be established and put all the points into set N. If not, label the point as noise. Then, do the same process on the points in the set. Repeat the process until all the points have been labeled to a cluster or noise.

What is needed for us to pay attention to is that the Eps and MinPts also need to be determined and it's not an easy work for we need to try a lot of values and do not know whether the choice is the best. Still we can do some preparation to make the process easier and faster. At first, we need to know how the result changes when we adjust the value of Eps and MinPts: if Eps is too large, all the points will be gathered into one group for the max radius is so large that all the points are in this range, while if the value is too small, the number of clusters will be so large that it's difficult to find relationship between the clusters. Also, if the MinPts is too large, maybe noise points will be named a cluster label which makes the result lose accuracy, while if the value is too small, the number of clusters will also be large. What's more, the result is related to the number of input data, that is to say, the minimum point number of a cluster should be determined based on that of the whole points set.

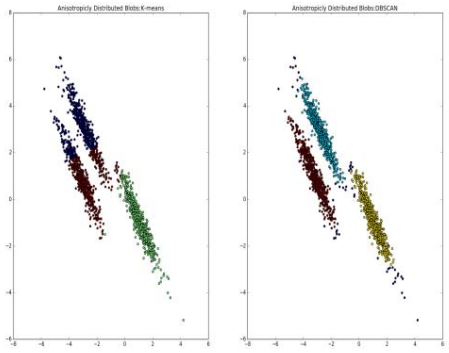

Figure 1: Comparing classification result using K-means and DBSCAN

Figure1 shows the clustering result of two algorithms. We can see that DBSCAN is a better way for it can remove the noise points and for this kind of distribution, if the classification does not depend on the density, the result is obvious wrong.

C. Principle Component Analysis(PCA)

PCA is always a good way of pretreating data to be the form we need especially in the condition where the data are high-dimensional and very difficult to do calculation in the following steps.

It uses an orthogonal transformation to convert a set of observations constructed from possibly correlated variables into a set of that with linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. How PCA works is firstly moving the center of data set to 0; then calculating the covariance matrix and getting k max eigenvalues of the matrix; at last, choosing the corresponding eigenvectors to form a matrix and then projecting samples to the vector.

III. EXPERIMENT

In oil industry, it is important to know the working condition of the oil wells and if we can detect the wells in bad working condition, it is possible for us to restore them to normal working in first time to ensure the productivity.

Nowadays in oil research area, it is convenient to measure the electrical parameters [9] and other operating parameters [10] of the pumping unit and do research on them. For example, the yield and efficiency show the productivity of the oil well; the oil pressure and casting pressure tells whether the well is in a normal working condition or not; the current, voltage, power of the well is also related to the working condition. Thus using these parameters, we can divide the wells into different types for easier management.

In consideration of characteristic of the data from oil well database and how the two algorithms operate, we need to do some pretreatment on the data.

The first step is loading data and doing pretreatment on the data. The data are extracted from the database of Dagang oil well and then filtered by program to get data of wells which are in the same block to be managed. If there is "None" value in the table, to make data be the same form, we set the "None" value to be zero. Then, the data have been put into a text and data of one well is shown in one row. After traverse the whole table, we put the well names into a string array and the rest data into a multi-dimensional matrix.

The second step is to do the dimensionality reduction. The reason is that if the data are two-dimensional, it is possible for us to show them in the picture. What's more, for the clustering algorithm, it is easy to calculate the distance between two-dimensional nodes. Thus, the data have been transformed to an array of n*2.

The third step is of course applying two algorithms on the data. For K-means clustering method, we set the number of clusters to be 3 after trying several different numbers and randomly choose 3 points to be the initial position of centroids. For DBSCAN clustering method, we should through experiment choose the maximum diameter and the minimum points of the same cluster. Then after the calculation, we show the result in two ways:

- In picture form: It is convenient for us to see the result: different colors represent different clusters, so it clearly shows which cluster the point belongs to but with the confusion of which well the point represents.
- In text form: this text will tell us all the well name of points in every cluster and solve the problem mentioned above.

As mentioned above, the accomplishment of the experiment is based on the libraries in Python: sklearn library can provide us with packaged algorithm, including K-means, DBSCAN and PCA; matplotlib library can give us a convenient way of plotting, for example, plotting the points in different clusters in different colors; numpy library does data processing quickly, like exchanging the data from list form to matrix form or on the contrary, adding elements to the existing matrix and so on, and some Python equations have requirement on the form of data thus data processing being helpful to analysis.

The process of choosing the parameters and reforming the data for later calculation is the main difficulty for me. There is only one parameter (number of clustering) to set in the program of K-means, while for the program of DBSCAN, we need to determine two parameters mentioned above. We firstly use the value in a program read before, but get too many clusters and dense point distribution. At this situation, what we can do is to enlarge the Eps or the MinPts. For the distribution is somehow dense, MinPts does not have much effect on the clustering result for even if the MinPts reduces, the points densely distributed still in the same cluster, while enlarging the Eps can contain more points with densely distribution into the same cluster. Thus, enlarging Eps can solve the problem above.

Data processing is also complicated, for using different types of data set, such as list or matrix, programming functions are different: the way we save the text of data is initializing an empty list firstly and then going through the text and dividing the information into the well name part and the data part. The loading process uses list form, so the function to put elements in the list is 'append', while in matrix form, we should use function 'insert' instead. Also, data in these two forms can be transformed to the other using functions from numpy library.

IV. Result and Comparison

Comparison of two clustering methods:

From the theory, we can find some differences between the algorithms:

- K-means is a clustering method related to the initial position of the centroids while DBSCAN depends on the density of the point distribution and has no initial settings. Seeing from Figure 1, we can find the differences: clustering according to density has the advantage of precise classification regardless of the shape of point distribution. In this way, DBSCAN is a better method. Also, the number of cluster for K-means also needs to be determined before the calculation and it may cause bad effect on the result. On the other hand, because of the property of randomly choosing initial cluster centroids in K-means, it may cause different classification results using different initial conditions. We can only get the locally optimal solution, and perhaps need to find the best by trail and error.
- If the densities of different clusters are much different from each other, the result won't be that good using DBSCAN for the least distance for clusters is all the same.

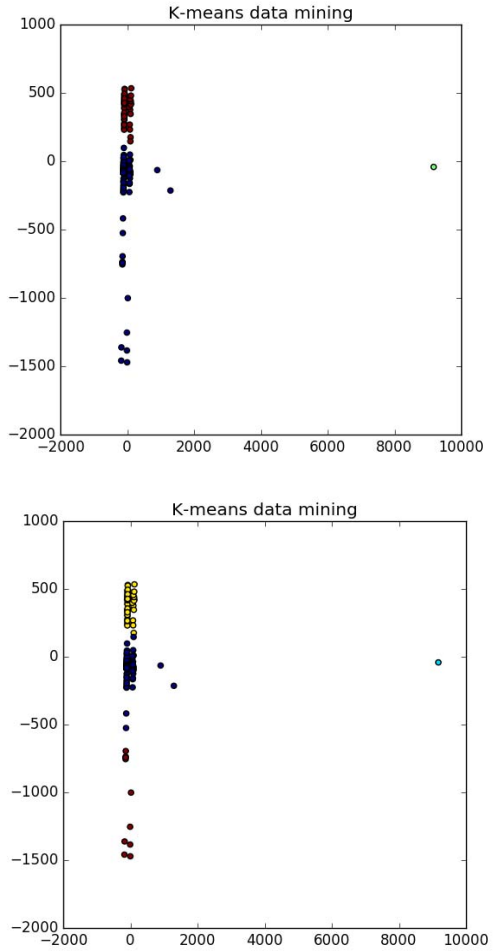

Fig. 2. Classification result with number of cluster 3 or 4

Taking a look at Figure2, we can see that if we divide points into 3 clusters, the noise point which is far from other points is divided into a cluster itself and thus the rest points

will be divided into only 2 clusters. For better comparison, we set the number of clusters to be 4.

Figure 3 shows the results of two algorithms, we can see that after removing the noise points, DBSCAN shows a clearer result. There are many points which are far away from the place where most points locate, and these points of course should not be divided to any group and they should be treated separately with different type of label--noise. Each cluster is far away from other clusters and points in the same cluster gather together so we can see the difference between clusters easily. The algorithm can help us divide the wells into different types and thus provide us a way of managing the oil field orderly.

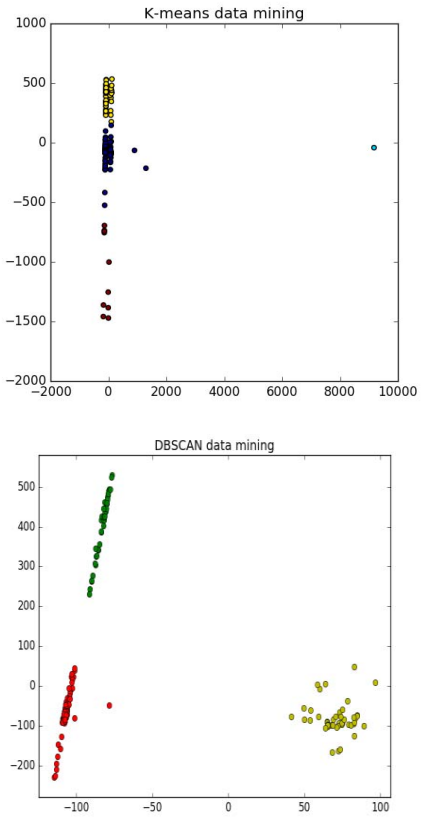

Fig.3. Classification result using K-means and DBSCAN

V. CONLUSION AND DISCUSSION

Our life can hardly do without oil--transportation, warming system, cooking and so on. Pumping unit is widely used in oil industry and production where the efficiency and working condition of machine is very important. When the machine is not working well, we need to discover and repair the breakdown quickly, and thus, management system based on classification is necessary. Through the process of clustering, the wells are divided into several clusters, and when we analyze specific wells to see what can be done to improve the efficiency, the adjustment operation may be also efficient for the wells in the same cluster. Clustering is hardly used in the oil production area nowadays, and it means a lot in practical application. After detailed analysis on the results, we see that DBSCAN clustering method has better classification result. To date, the clustering method is only used to classify the different types of oil pumping unit or identify the quality of oil.

There is still a lot to be done in the future research. For the consideration of minimizing our calculation time especially important when the dataset is large and difficult to analyze, we can use parallel processing. We know that after PCA dimensionality reduction, the axis does not have any meaning for the vector contains the mixing information of all dimensions. For better understanding, we can divide the data into two parts: efficiency group and productivity group and do PCA processing separately. The two group only contains parameters which are only related to efficiency or productivity. Thus, the axis in the picture has clear definition and it makes sense in production.

ACKNOWLEDGMENT

The author wants to thank Shiwei Ren and Weijiang Wang for their useful advice and the members of Lab for their help.

REFERENCES

[1] Sun Jigui, Liu Jie, Zhao Lianyu. Clustering Algorithms Research[J]. Journal of Software, 2008, 19(01):48-61.(in Chinese)

[2] Hartigan J A, Wong M A. A K-means clustering algorithm.[J]. Applied Statistics, 2013, 28(1):100-108.

[3] Yang Shanlin, Li Yongsen, Hu Xiaoxuan, Pan Ruoyu. Optimazition Study on k Value of K-means Algorithm[J]. Systems Engineering & Theory Practice. (in Chinese)

[4] Yu Jian , Chen Qiansheng.The range of optimal class number of fuzzy cluster[J] .Science of China(series E), 2002, 32(2):274-280. (in Chinese)

[5] Fan Jiulun, Pei Jihong , Xie Weixin.Cluster validity function:Entropy formula[J] .Fuzzy Systems and Mathematics, 1998, 12(3):68-74. (in Chinese)

[6] Jiawei Han , Micheline Kamber.Data Mining Concepts and Techniques[M] .Fan Ming , Meng Xiaofeng, et al.Beijing:China Machine Press, 2001. (in Chinese)

[7] Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M., eds. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231. ISBN 1-57735-004-9. CiteSeerX: 10.1.1.121.9220.

[8] Rong Qiusheng, Yan Junbiao, Guo Guoqiang. Research and Implementation of Clustering Algorithm Based on DBSCAN[J]. Computer Applications , 2004, 24(04):45-46. (in Chinese)

[9] Chen Shi, Wang Haiwen. Diagnosis Technology Using Electricity Parameter Method for Production Well Adopting Progressive Cavity Pump[J]. Oil Field Equipment, 2007, 36(2):53-55. (in Chinese)

[10] Nie Feipeng, Ma Ying, Zhang Xuemei,et al. New Way of Working Condition Diagonosis on Spiral wells[J]. Fault –Block Oil & Gas Field, 2007, 14(06):76-77.