K-means and Support Vector Machine in Electric Power Company Benchmarking Management

Hong-qing Zhang

School of Economics and Management, North China Electric Power University, Beijing, China

Abstract - In the electric power company benchmarking management, implementing classification of the enterprise, the clustering algorithm can set up the model enterprise**.** It's very important for the benchmarking management in the electric power company**.** K-means, as unsupervised learning algorithm, is suitable for processing great sample data, while support vector machine(SVM), as supervised learning algorithm, needs a small number of training samples and is able to obtain the higher classification accuracy**.** Therefore, the paper presented a classification method based on the combination of SVM and K-means**.** Using K-means clustered index data first, and then chose some samples which were close to each cluster center as study samples to training SVM classifier and classified all the index data with SVM classifier**.** Consequently, illustration showed that K-means combined with SVM had higher accuracy than K-means, which testified the validity of it**.**

Index Terms - K-means; SVM; Benchmarking Management; Electric Power Company

1. Introduction

Compared with the model enterprise, enterprises can find their own advantages and disadvantages. It not only can improve the development of enterprise, but also can give full play to comparative advantage in the market of the enterprise and enhance enterprise market competence. Eventually, how to choose the model enterprise in the benchmarking management is crucial. At present, scholars in types of organization proposed all kinds of model methods.

In the nonelectric power industry, Ref. [1] established the TOPSIS enterprise benchmarking evaluation model and the model was applied to the comprehensive evaluation of benchmarking management in six selected international petroleum enterprises. This established TOPSIS comprehensive evaluation model and accompanying benchmarking management system are not only to evaluate operating performance and management level, but also achieve monitoring during the whole process which is of great significance for enterprises to improve value creation capacity and international competitiveness. Ref. [2] introduced the cross-evaluation into the improved-DEA model to distinguish the effective decision-making units and to select the benchmarking enterprises for CNPC.

In the Electric Power Industry, Ref. [3] constructed the performance evaluation index system for grid enterprises, and established a comprehensive performance evaluation model by combing analytic hierarchy process and fuzzy comprehensive evaluation method, according to the characteristics of grid enterprise. And it verifies the feasibility of the above model through an example. Ref. [4] built gray models of making the leading enterprise and classifying the enterprise as well as comprehensive evaluation of indexes separately in connection with the gray characteristic of the electric benchmarking comprehensive evaluation, and the model has validity and usability. Ref. [5] showed that the choice should be based on the method of individual advantage identification, and gave detailed procedures of implementation, and it is indicated that the model is scientific and objective. Ref. [6] made K-means clustering algorithm apply to the indicators comparison of power enterprises. And through the example, it is showed that the proposed algorithm is effective. Ref. [7] came up with the applications of combination of gravitational search and K-means algorithm in the benchmarking management. Besides, it is proved that the algorithm has validity and feasibility.

Although K-means algorithm is simple and efficient, it make cluster results produce errors when sample data used is not numerous enough. While, SVM needs a small amount of training samples and can has the high classification accuracy. As a consequence, the combination of K-means and SVM classification was proposed. Not only can it increase accuracy of classification, but also solve the problem that accuracy of K-means is not high.

2. Theories of K-means and SVM

A. K-means

K-means, as unsupervised learning algorithm, is good at disposing big-sample data. It is needed to set k-clustering number, and choose the initial k clustering centers. According to the minimum distance criterion, the data is distributed to the k classes. Then, calculating the average distance of the data in each class and the initial clustering center can achieve new k clustering centers. If the new clustering centers and clustering centers in the last iteration are the same or the new clustering centers are less than the convergence criteria which is defined subjectively, the clustering is over Otherwise, the algorithm is getting to the next iteration. Selecting the initial clustering centers accurately will greatly reduce the iteration steps because the result of the algorithm relies on the choice of initial clustering centers.

B. SVM

SVM is a new machine learning method based on statistical learning theory and it's worth knowing that SVM is used all over the place to solving classification problems of small samples ,high dimension data and nonlinear. SVM is a classification model of two types. The basic model of SVM is

defined as largest interval linear classifier in the feature space and the learning objective of SVM is constructing a hyper plane as decision plane in the high-dimensional space, which makes the largest classification interval between two classes of data.

3. The Classification Method based on the Combination of SVM and K-Means

The classification process of the K-means which is an unsupervised classification is convenient, but the classification results are not good enough. However, SVM, a supervised classification can obtain high classification accuracy through training several sample data, but it needs manual identified samples for training, which makes the process relatively cumbersome. According to the advantages and disadvantages of each method, the paper presented a classification method in the electric power company benchmarking management based on the combination of SVM and K-means. This method avoided the shortcoming of unsupervised classification and eliminated the cumbersome process of manual identifying samples of SVM.

The major idea of the model is K-means was used to cluster original index data of enterprises firstly, and then according to the number and sparse degree of points in each class, some points as labeled samples were chosen to train SVM, at last SVM was utilized to reclassify original index data of enterprises.

Specific algorithm was described as follows:

A. Initial Cluster

$D(t)=[d_1(t),d_2(t),\dots ,d_n(t)]^T$ were the input samples The input samples were made to close to their clustering center by K-means, leading to create k clusters.

B. The Choice of Training Samples

Choosing the training samples, which were created by K-means and close to each cluster center, as training data of SVM.

C. Classifying

All the input samples were classified with SVM, getting the new sample classification.

4. Analysis of the Example

Some electric power enterprises in China were classified from the aspects of quality of power supply through making use of the combination of SVM and K-means. The indexes of power supply quality in China are mainly consisted of urban comprehensive power supply voltage percent of pass, rural comprehensive power supply voltage percent of pass, urban power supply reliability and rural power supply reliability. The index data related to the quality of power supply of all of province electric power companies in 2011 and corresponding classes were shown in Table 1. Samples of the 20 enterprises were divided into the first class, the second class and the third class according to the comprehensive ranking.

Before using K-means, the data were needed to the positive treatment and make indexes being dimensionless. The methods of positive treatment are negated and so on. The methods of making indexes being dimensionless are extremum method, difference artithmetic and so on. Since all the indexes in Table I are positive and dimensionless, they didn't need to preprocess data and were straightforward to use K-means.

TABLE I The Index Data Related to the Quality of Power Supply of All of Province Electric Power Companies in 2011 and Corresponding Classes

Enterprise	Urban comprehensive power supply voltage qualification rate	Rural comprehensive power supply voltage qualification rate	Urban power supply reliability	Rural power supply reliability	Class
1	99.86	99.62	99.981	99.93	1
2	99.45	98.11	99.931	99.762	1
3	97.16	97.16	99.911	99.765	3
4	99.48	98.81	99.971	99.94	1
5	99.81	98.92	99.97	99.907	1
6	96.5	96.4	99.954	99.809	3
7	99.69	99.6	99.925	99.626	1
8	99.3	94.88	99.894	99.761	3
9	99.95	99.83	99.953	99.879	1
10	99.95	99.86	99.973	99.806	1
11	96.68	96.65	99.947	99.769	3
12	99.81	97.46	99.983	99.94	1
13	99.79	97.1	99.938	99.739	2
14	99.73	96.89	99.956	99.857	2
15	99.82	97.83	99.945	99.768	2
16	98.24	96.57	99.943	99.781	2
17	99.71	99.78	99.94	99.684	1
18	99.88	97.31	99.918	99.79	2
19	99.94	96.98	99.929	99.745	2
20	98.93	98.29	99.929	99.887	1

Let the number of initial clustering is 3, the indexes of 20 electric power enterprises were clustered by using SPSS. The results of clustering and the distance from each data to clustering center were shown in Table II.

TABLE II The Results of Clustering and the Distance from Each Data to Clustering Centers

Enterprise	Clustering class	The distance from each data to clustering center	Enterprise	Clustering class	The distance from each data to clustering center
1	1	0.59937	11	3	0.4674
2	1	0.98078	12	2	0.70787
3	3	0.46634	13	2	0.34006
4	1	0.33806	14	2	0.13264
5	1	0.22393	15	2	1.06358
6	3	0.70998	16	3	1.10212
7	1	0.56989	17	1	0.7291
8	2	1.94189	18	2	0.55795
9	1	0.81857	19	2	0.29518
10	1	0.84491	20	1	1.07052

Chose enterprise number 3,4,5,14 and 19 which were closest to the clustering center as training samples of SVM classifier as independent variables matrix 'index', clustering class as dependent variables matrix 'label' .They were trained

in SVM classifier by using Matlab. For the number of clustering class is 3, '-1' was on behalf of the first class, '0' was on behalf of the second class and '1' was on behalf of the third class. After training samples, the 20 samples as test suite were classified by SVM classifier.

Parts of the codes were as follows:

```
index=[97.16,97.16,99.911,99.765;
99.48,98.81,99.971,99.94;
99.81,98.92,99.97,99.907;
99.73,96.89,99.956,99.857;
99.94,96.98,99.929,99.745;];
label=[1;-1;-1;0;0];
[x,ps]=mapminmax(index');
model=svmtrain(label,x',' -s 1 -c 16 -t 0 -g 1 -r 5 -d 3');
label_truth=[1;1;3;1;1;3;1;3;1;1;3;1;2;2;2;2;1;2;2;1];
xtest=[
99.86, 99.62, 99.981, 99.93;
99.45, 98.11, 99.931, 99.762;
97.16, 97.16, 99.911, 99.765;
99.48, 98.81, 99.971, 99.94;
99.81, 98.92, 99.97, 99.907;
96.5, 96.4, 99.954, 99.809;
99.69, 99.6, 99.925, 99.626;
99.3, 94.88, 99.894, 99.761;
99.95, 99.83, 99.953, 99.879;
99.95, 99.86, 99.973, 99.806;
96.68, 96.65, 99.947, 99.769;
99.81, 97.46, 99.983, 99.94;
99.79, 97.1, 99.938, 99.739;
99.73, 96.89, 99.956, 99.857;
99.82, 97.83, 99.945, 99.768;
98.24, 96.57, 99.943, 99.781;
99.71, 99.78, 99.94, 99.684;
99.88, 97.31, 99.918, 99.79;
99.94, 96.98, 99.929, 99.745;
98.93, 98.29, 99.929, 99.887;
];
[xtest1,ps1]=mapminmax(xtest');
[predict_label]=svmpredict(label_truth,xtest1',model)
[predict_label]=svmpredict(label_truth,xtest1',model)
Accuracy = 90% (18/20) (classification)
predict_label=[1;1;3;1;1;3;1;2;1;1;3;1;2;2;1;2;1;2;2;1;]
```

The results of SVM classifier were indicated in Table III.

As shown in Table IV, the accuracy of the trained model was 90%, and among them, 18 classifications of the enterprises were correct, two of them were wrong. The clustering precision was 75% when using K-means only, and among them, 3 classifications of the enterprises were wrong. The example proved that the algorithm based on the combination of K-means and support vector machine had higher accuracy.

TABLE III The Results of SVM Classifier

Enterprise Clustering class	Clustering class The distance from each data to clustering center	Enterprise	The distance from each data to clustering center
1	1	11	3
2	1	12	1
3	3	13	2
4	1	14	2
5	1	15	1
6	3	16	2
7	1	17	1
8	2	18	2
9	1	19	2
10	1	20	1

TABLE IV Accuracy of Two Arithmetics

Arithmetic	K-means	K-means and SVM
Accuracy%	85%（17/20）	90%（18/20）

5. Conclusion

Using the combination of K-means and SVM to get classification accuracy of 20 power companies is higher than K-means. Thus, the combination of K-means and SVM uses a few training samples to train SVM classifier and gets high accuracy, solving the problem that the classification accuracy of K-means is good enough.

K-means combined with SVM algorithm is attempted to classify in the electric power company benchmarking management and it has certain feasibility. However, the algorithm is only used for classification and it's not enough for the analysis model of electric power enterprise benchmarking management. In the future, researchers need to propose more effective and feasible model of electric power company benchmarking management according to the electric power enterprise development, so that they can provide new ideas for the benchmarking management in electric power enterprise.

References

[1] X. Xu, J. Liu, and Z. Wang, "Enterprise benchmarking evaluation model based on entropy- weighting TOPSIS method and empirical research," in *Journal of Information*, vol. I, Press, 2011, pp. 78-92.

[2] H. Ding, Y. Cao and Y. Luo, "Study of CNPC association benchmarking based on DEA and cross-evaluation," in *Journal of Gansu Sciences*, vol. II, Press, 2012, pp. 151-154.

[3] W. He, F. Zhong, and Y. Chang, "Study on investment performance evaluation of grid enterprise," in *Technical Economy*, vol. I, Press, 2011, pp. 78-84.

[4] J. Yang, Research on electric benchmarking model based on grey decision making, unpublished master dissertation, North China Electric Power University, Beijing, China, 2008.

[5] B. Wang, A method to choose surveyor's pole based on individual advantage identification and its application in standard target system, unpublished master dissertation, Northeastern University, Shenyang, China, 2007.

[6] W. Xue et al, "Application of K-means algorithm in indicators comparison of power industry" in Hydroelectric Energy, vol. VI, Press, 2012, pp. 240-242+13.

[7] X. Liu, Research on benchmarking for grid enterprises based on data mining, unpublished master dissertation, North China Electric Power University, Beijing, China, 2014.

[8] N. Shrivastava, S. Sharma, and K. Chauan, “Efficiency assessment and benchmarking of thermal power plants in India,” in Energy Policy, vol. XL, Press, 2012, pp. 159-176.
[9] A. Haney, and M. Pollitt, “Exploring the determinants of best practice benchmarking in electricity network regulation,” in Energy Policy, vol. XXXIX, Press 2011, pp. 7739-7746.
[10] T. Kuosmanen, A. Saastamoinen, and T.Sipiläiene, “What is the best practice for benchmarking regulation of electricity distribution? Comparison of DEA, SFA and StoNED methods,” in Energy Policy, vol. LXI, Press 2013, pp. 740-750.
[11] D. Chan et al, “Energy efficiency benchmarking of energy-intensive industries in Taiwan,” in Energy Conversion and Management, vol. LXXVII, Press 2014, pp. 216-220.
[12] W. Chung, “Review of building energy-use performance benchmarking methodologies,” in Applied Energy, vol. LXXXVIII, Press 2011, pp. 1470-1479.