Paper 251-26

Quality Design based on SAS/EM®

Wenming Chen, Baosteel Research Institute, Shanghai, 201900, China
Dongping Su, Baosteel Research Institute , Shanghai, 201900, China
Jiansheng Feng, Baosteel Research Institute, Shanghai, 201900, China

ABSTRACT

During the past two decades, our company Baosteel (the Baoshan Iron and Steel Corporation, Shanghai, China) has accumulated enormous data and information. Much knowledge, such as relationship, rules and patterns among these raw data, is hidden. It was very hard to reveal completely the knowledge since no advanced tools existed. Here we thank SAS Institute for providing us powerful tools to solve these problems. A lot of useful information in our project has been mined after SAS Enterprise Miner tools are used. This paper focuses on using SAS software to control and improve the quality of products in Baosteel. According to the results of mining, the engineers can make their decisions more accurately and easily and change their traditional methods of improving quality by qualitative or by testing.

The SAS products include SAS/BASE®, SAS/CONNECT®, SAS/ACCESS®, SAS/IML®, SAS/STAT®, and the like for windows™ 95. Besides these basic tools, our more useful and efficient tool is SAS/Enterprise Miner 3.0® for windows™ 95.

The skill level for intended audience: beginner and above.

1. INTRODUCTION

It is very hard for quality engineers to design and control product quality in the manufacturing enterprises with multi-process. Sometimes, some production parameters must be adjusted in order to improve the quality of products. However, how can they do it? One possible case is to adjust these parameters for the engineers according to their experiences. The other possible case is to do some experiments for the researchers in their laboratories, but the results of experiment are different from that of real production because the situation in the laboratory is not as same as that of the large-scale manufacture. On the other hand, modern manufacturing enterprises have accumulated plenty of production data since various computer systems are applied. For example, Baosteel has accumulated gigabytes of production data since it was put into production in 1980s. As we know, data are asset, but they are not utilized enough. Therefore it is worth to apply KDD (Knowledge Discovery in Data) approach to analyze the practical production data to benefit the design and the control of product quality.

2. SEVERAL CONCEPTS AND KDD PROCEDURES

2.1 DATA PREPARATION

Data preparation provides data mining with the appropriate data. At first, the data distributed in different working processes and quality tests are integrated into a data warehouse. In this paper, data mart[1]---the small local data warehouse---is adopted. Before the data mining procedure, the data need doing some preprocessing, such as filtering, sampling, missed value processing, variables selection, coding, and so on. The SAS/BASE®, SAS/CONNECT® and SAS/ACCESS® are useful in this procedure.

2.2 DATA MINING

Data mining is the most important process of the knowledge acquisition. There are many methods of data mining, such as statistical methods, cluster analysis, pattern recognition, decision tree, association rule, artificial neural network, genetic algorithm, rough set theory, visualization technology, and so on.[2]

After comprehensive comparison, the cluster analysis is chosen as our primary data mining method because it is an unsupervised method and appropriate to our application. Clustering algorithms segment the data into groups of records, or clusters, that have similar characteristics. The result of cluster analysis can be interpreted into qualitative knowledge, adopting the format of if-then rule. They can tell engineers what are the main factors that affect the quality of product and the possible directions when someone wants to adjust the production parameters. The qualitative knowledge can be upgraded into quantitative knowledge if some additional works are done. These additional works include modeling with regression analysis or artificial neural network. By using this knowledge, quality engineers can design and control product quality more easily, effectively and efficiently than ever.

The SAS/Enterprise Miner 3.0®(i.e. SAS/EM®) is a useful data-mining tool, in which analytical methods integrated with powerful visualizations make one fulfill the work efficiently. SAS/EM® includes six kinds of tools---sample tools, explore tools, modify tools, model tools, assess tools and utility tools, and it can directly implement a lot of data mining methods aimed at various applied fields. There still are some methods not included in SAS/EM®, just like genetic algorithm and rough set theory, but they can also be implemented with SAS programming indirectly.

2.3 TWO PROBLEMS OF CLUSTER ANALYSIS

There are two problems needed to solve in the cluster analysis,

one is what is the most appropriate number of clusters, the other is how to segment the data into clusters when the number of clusters is given.

For the second problem, many clustering algorithms can be used. In SAS/STAT®[3], there are 11 kinds of hierarchical clustering algorithms involved in CLUSTER procedure step and a dynamic clustering algorithm involved in FASTCLUS procedure step, but the clustering algorithms based on artificial neural network are absent. Therefore we use SAS/IML®[4] to implement clustering algorithms based on artificial neural network. There are less clustering algorithms to be chosen in the Clustering Node of SAS/EM®, but the algorithms in common use are held. In SAS/EM®, an additional SOM Node can play a role of clustering algorithms based on artificial neural network---the SOFM network, but it is a pity that the ART network families are still lacking.

The first problem is more difficult than the second one. There are no satisfactory methods for determining the number of population clusters for any type of cluster analysis. Some researchers, such as Arnold, Sarle, Milligan, Cooper, Wong, made their contributions to this problem respectively. In SAS/EM®, a cubic clustering criterion is used to determine the optimal number of clusters if the user chooses the automatic cluster option. However, it is inclined to form fewer clusters in practice. In this paper, a kind of error square sum criterion is adopted.

Let us consider clustering n observations into c clusters. Note the

$$J = \sum_{i=1}^{c} \sum_{x \in G_i} \left\| x - \bar{x}_i \right\|^2 \qquad \text{...(1)}$$

measures the square sum of error caused by replacing c clusters $G_1, \cdots, G_c$ by their centers of clusters $\bar{x}_1, \cdots, \bar{x}_c$.

Given c, the J value is expected as minimal as possible. Obviously if $c = 1$, all observations are involved in one cluster, then J equals the product of the variance of the observations and the number of the observations. Let $c = n$, each observation is put into one cluster, then $J = 0$. Generally speaking, the J value will decrease along with the increase of the c value, so we can plot a $J - c$ curve, and the c value corresponding to the inflexion of this curve can be taken for the most appropriate number of clusters. For instance, the best number of clusters is 6 in Figure 1.

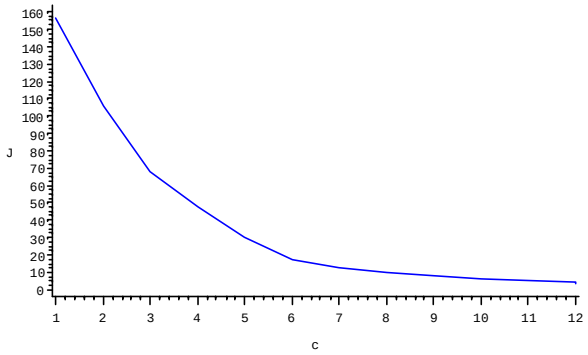

Figure 1 an example of applying $J - c$ curve to determine the most appropriate number of clusters

Some of these clustering algorithms are more effective than others in our special data environments. They are the average linkage algorithm, the centroid algorithm, the Ward's minimum-variance algorithm, the k-mean algorithm and the clustering algorithms based on artificial neural network. However, it is hard to say which one is the best for various cases.

2.4 ONE OF THE MODERN QUALITY CONTROL CONCEPT

In modern quality control, C_{pk} (Capability Index) is an important concept. For the product quality with upper and lower constraints, C_{pk} is defined as:

$$C_{pk} = \frac{S_U - S_L}{6\sigma} \qquad \text{...(2)}$$

For the product quality with only upper constraint, C_{pk} is defined as:

$$C_{pk} = \frac{S_U - m}{3\sigma} \qquad \text{...(3)}$$

For the product quality with only lower constraint, C_{pk} is defined as:

$$C_{pk} = \frac{m - S_L}{3\sigma} \qquad \text{...(4)}$$

Notation: S_L and S_U stand for the lower constraint and the upper constraint respectively, m and σ stand for the mean and the standard deviation of samples respectively.

It had better to control C_{pk} within a range. The product quality can not be guaranteed if C_{pk} is too low. Some resources are wasted unnecessarily and the cost gets too high if C_{pk} is too high. Our target is to attain the first-class treatment, which requires $1.33 < C_{pk} \le 1.67$.

3. AN APPLICATION TO QUALITY IMPROVEMENT

3.1 THE ORIGINAL DATA FEATURE

The data mart includes more than 20,000 observations and 35 variables, among which three variables are KOVs (Key Output Variable) and the others are KIVs (Key Input Variable). A subject dataset for further analysis is drawn out according to some special rules, which still includes about 5000 observations.

All KOVs of the special product are needed to meet their lower constraints respectively, so C_{pk} is calculated by the formula (4). The original means, standard deviations, lower constraints and C_{pk} of the three KOVs are showed in Table 1. Now the problem is to increase C_{pk} of KOV2 and KOV3 and decrease C_{pk} of KOV1.

	KOV1	KOV2	KOV3
MEAN	35.92	392.77	531.91
STD	2.15	18.69	15.78
S_L	22	350	480
C_{pk}	2.16	0.76	1.10

Table 1 the original means, standard deviations, lower constraints and C_{pk} of the three KOVs

3.2 THE SUPERIOR CLUSTERS AND THE INFERIOR CLUSTERS

According to the formula (4), the means of KOVs ought to be increased and the standard deviations of KOVs ought to be decreased in order to improve C_{pk} of KOVs. Therefore the clusters with large means and small standard deviations of KOVs are considered better than the clusters with small means and large standard deviations of KOVs. In this paper, the formers are called the superior clusters and the latters are called the inferior clusters. The parameter adjustment can be determined among the production parameters of superior clusters. Of course, the adjustment must be consistent with the opinions of the product experts before it can be applied in practice.

3.3 THE SAS/EM® DIAGRAM

The diagram of data mining is shown in Figure 2. As we mentioned previously, the cluster analysis is chosen as our primary data mining method, the result of clustering helps engineers to make their decisions. In addition, a Neural Network Node is used to simulate the system and predict the possible results after the production parameter adjustment. The other nodes are auxiliary.

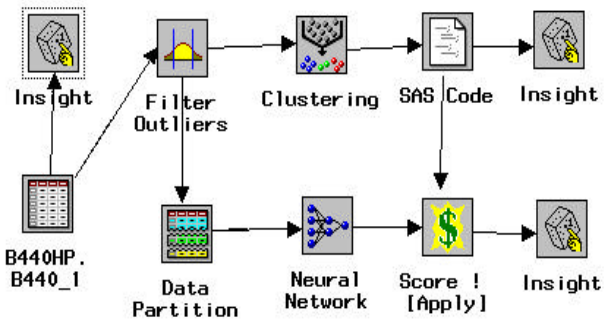

Figure 2 the diagram of data miming in our application

3.4 THE RESULT OF CLUSTERING

At first, the appropriate number of clusters is needed to determine. A SAS/IML®[4] program is written to do it. According to the error square sum criterion, the most appropriate number of clusters is 16. Then we specify the number of clusters as 16 in the Clustering Node. The result of clustering is shown in Figure 3 and Figure 4.

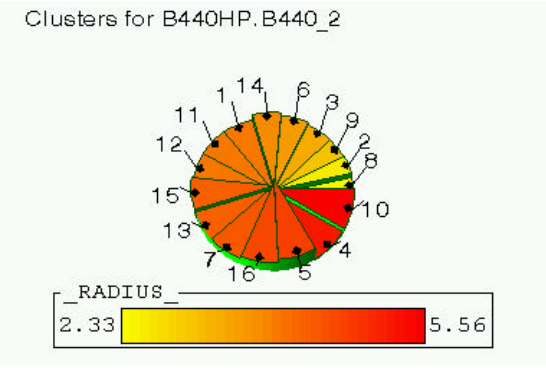

Figure 3 the pie chart of clustering result

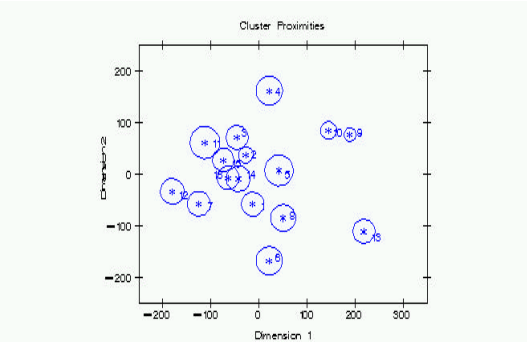

Figure 4 the distance chart of clustering result

3.5 THE IMPROVED C_{pk}

After the KDD approach is applied, C_{pk} of the three KOVs improves, as the result showed in Table 2.

	KOV1	KOV2	KOV3
MEAN	35.09	404.28	542.60
STD	2.07	14.70	12.17
C_{pk}	2.11	1.23	1.71

Table 2 the improved means, standard deviations and C_{pk} of the three KOVs

The result is encouraging, but there still are some imperfections. C_{pk} of KOV2 has been lower than 1.33 yet. We think it due to the fact that the original mean of KOV2 is not large enough and the original standard deviation of KOV2 is large. More researches are needed in order to make the better performance.

CONCLUSION

It is proved effective to help quality engineers to design and control product quality by using KDD approach. SAS/EM® is a very powerful KDD tool for both industry and commerce, it can accelerate our research project. We believe there are more KDD applications in our manufacturing enterprise in the future.

REFERENCE

[1] W.H.Inmon. Building the Data Warehouse, 2nd ed., John Wiley & Sons, 1996.

[2] P.Adriaans and D.Zantinge, Data Mining, Addison-Wesley, 1996.

[3] SAS Institute Inc. SAS/STAT® User's Guide, Version 6, 4th ed., Vol. 1, 1990.

[4] SAS Institute Inc. SAS/IML® Software: Usage and Reference, Version 6, 1990.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Please contact the author at

Chen Wenming

Automation institute, Baosteel Technology Center,

Baoshan, Shanghai, 201900,China

E-mail: wmchen@usa.net