Analyzing Operation Deviation in the Deasphalting Process Using Multivariate Statistics Analysis Method

Joo-Hwang Park*, Jong-Soo Kim**, Tai-Suk Kim***

ABSTRACT

In the case of system like MES, various sensors collect the data in real time and save it as a big data to monitor the process. However, if there is big data mining in distributed computing system, whole processing process can be improved. In this paper, system to analyze the cause of operation deviation was built using the big data which has been collected from deasphalting process at the two different plants. By applying multivariate statistical analysis to the big data which has been collected through MES(Manufacturing Execution System), main cause of operation deviation was analyzed. We present the example of analyzing the operation deviation of deasphalting process using the big data which collected from MES by using multivariate statistics analysis method. As a result of regression analysis of the forward stepwise method, regression equation has been found which can explain 52% increase of performance compare to existing model. Through this suggested method, the existing petrochemical process can be replaced which is manual analysis method and has the risk of being subjective according to the tester. The new method can provide the objective analysis method based on numbers and statistic.

Key words: Deasphalting Process, Big Data, Operation Deviation, Multivariate, Statistics Analysis

1. INTRODUCTION

System is the source of big data includes basic system of corporate management system ERP (Enterprise Resource Planning), MES(Manufacturing Execution System) which focuses on production automation control, company-wide plant operation management, environmental management to support the PIS(Plant Information System) and so on[1–3].

Among the computing system which is to monitor the process of crude oil refining, by using multivariate statistical process control which is suggested to improve the process, the method for reducing the operation deviation in deasphalting process can be decided quickly.

By using statistical analysis that utilizes big data system to find out optimal condition of various equipment at the petrochemical process, it can replace existing manual analysis method which can shorten the analysis time and save the cost and business productivity and management activity can be improved using the big data from distributed computing system that business possess[4–5].

2. RELATED RESEARCH

A variety of statistical method can be applied to analyze deasphalting process which is one of the
Analyzing Operation Deviation in the Deasphalting Process Using Multivariate Statistics Analysis Method

Various chemical compounds can be used as solvent before the process. And after the process, quantity and quality of the output can be different depending on the solvent which have been used [6–9].

Through these characteristics of the process, crude oil (Feed) can be seen as dependent variable which is the main ingredient, and the asphalt can be seen as independent variable which is the subsidiary materials. Therefore, it is possible to apply multiple regression analysis from these the statistical models.

Input variables before the petrochemical process are quantity of oil before the process, specific gravity, temperature, viscosity, and the type of solvent. Reactor variables can be defined as a Process variable; Output variables after the process are the quantity of oil after the process, specific gravity, temperature, viscosity, and the quantity of by-product.

In order to analyze the desasphalting process, subsidiary materials AR, VR, and Solvent (FIC1307) were added and Fig. 1 shows the asphalt extracting process using process variables—INPUT, PROCESS, OUTPUT.

To analyze operation deviation, various variables should be taken into consideration such as other working condition apart from Input, Process, and Output variables.

The multiple regression equation, which is used to apply statistical model in process analysis, generally defined as equation (1) [10–11].

\[ Y = \alpha + \beta_1 X_1 + \cdots + \beta_n X_n + \epsilon \] (1)

There are many ways to estimate the multiple regression equation that represents the characteristics of the process. Common methods are Simultaneous Input Method and Step input method. Besides these, backward method is the way to draw the regression equation by eliminate the less important variables sequentially after inputting all the independent variables to the regression formula.

The purpose of statistical analysis is to interpret interdependence and dependence relationships of different (various) variables in multivariate statistical analysis. And this statistical analysis can be used to interpret manufacturing process.

3. SYSTEM DESIGN

Fig. 2 shows data processing system of oil corre-
MSPC (Multivariate Statistics Process Control) method is used which is the new method that overcomes existing SPC (Statistics Process Control)’s limitation.

Fig. 3 shows the advantages of MSPC compared with SPC.

Regarding the analysis of running condition, it is founded that suggested MSPC-Chart is easier to distinguish than SPC process chart that shows abnormal running condition.

To compare the cluster such as operation deviation at the parallel processing system based on MPI (Message Passing Interface), Flow chart of applicable K-Means Clustering algorithm is schematized at Fig. 4.

K-Means Clustering Algorithm is Non-hierarchical clustering method and is formed the cluster by allocating each individuals to the closest central points.

Based on the monitoring result according to the change of Clustering state, Parallel processing code for K-Means Clustering can be designed such an example code as table 1.

With regards to design the application, K-Means algorithm was optimized at first, and then Distributed Parallel Processing method was used.

About 1.8 million cases (200 MB data) from different business field that uses K-Means algorithm were used to test the performance of algorithm optimization and Distributed Parallel Processing. Fig. 5 represents the results.

When K-Means algorithm was applied, it took 102 seconds to process the data. After the algorithm optimization, calculation time was shortened to 91 seconds which is 1.12 times performance improvement.

When Distributed Parallel Processing was applied, it took 102 seconds to process the data. After the algorithm optimization, calculation time was shortened to 91 seconds which is 1.12 times performance improvement.

### Table 1. An example code of a K-Means Clustering

```plaintext
for t = d+1:d+bufferLength
    clusterID(:, t) = autodbscan_corr(data(:, t-d:t));
    clusterDist(t) = distCluster_corr(data(:, t-d:t), clusterID(:, t));
    end

for t = d+1:size(data, 2)
    currentT = t+bufferLength;
    clusterID(:, currentT) = autodbscan_corr(data(:, currentT-d:currentT));
    clusterDist(currentT) = distCluster_corr(data(:, currentT-d:currentT), clusterID(:, currentT));
    mergeScore(currentT) = scoring_merge(clusterID(:, currentT-d:currentT), clusterDist(t:currentT));
    splitScore(t) = scoring_split(clusterID(:, currentT-d:currentT), clusterDist(t:currentT));
    numcluster_1 = max(clusterID(:, currentT)) + 1;
    if(numcluster_1 >1)
        plotMerge
    end
    numcluster_2 = max(clusterID(:, t)) + 1;
    if(numcluster_2 >1)
        plotSplit
    end
end
```

... //The implementation is omitted.
Analyze Operation Deviation in the Deasphalting Process Using Multivariate Statistics Analysis Method

4. OPERATION DEVIATION ANALYSIS

At the deasphalting process, as Figure 1 shows, to analyze operation deviation of solvent deasphalting process, regression equation for searching the factors that affects DAO YIELD can be simplified as equation (2).

\[
\text{DAO Yield} = \frac{\text{SDA1SDFIC1309}}{\text{SDA1SDFIC1301} + \text{SDA1SDFIC1302}} = \text{Function(Main arguments in all variables)}
\]  
(2)

The data status to analyze the operation deviation was collected between 09:01 June 24, 2010 to 10:45 May 22, 2013. SDA variables were collected per each 253 variables, and in total, Row 1530,820/minute unit. Table 2 shows the name of tag of SDA variables.

STATS for effective analysis of a non-zero value is deleted, then delete STATS Field, DAO Yield delete more than one value, *. Txt 10MB or less, due to the problem of matching with other data, except for values less than or equal to 0 in the variable value analysis (FIC1301, FIC1302, FIC1307, FIC1309, PI1302, TIC1303, VDU2A1107) that was deleted.

In the final analysis, the data is July 6, 2010 – May 22, 2013 9:32 10:40 a variable period of about 254 Row 1,095,506 / were prepared in minutes, and 253 variables Based on the statistical analysis was performed.

By analyzing the correlation between adjustable parameters mainly based on DAO yield, it is founded that Flow-related variable has high correlation as it shows in Fig. 6.

In analysis, 12 Adjustment variables were found that have high correlation coefficient. FIC1301 showed as variables that has the biggest correlation coefficient which is about -0.63 from the whole adjustable variables.

After choosing 81 adjustable variables in refer-
ence to the TAG, PCA analysis was performed and specific group was visualized by using K-Means analysis as it shows at Fig. 7.

The research was carried out to form two groups to compare first operation condition and second operation condition.

As information of variable was not available at the current stage, with the utilization of T-Square chart, outliers for the 81 adjustable variables were removed by 99% standard.

Operation deviation was analyzed by using adjustment variables which outliers were removed. To demonstrate the variable's contribution to main component at the PCA analysis, the Loading Plot chart was used. At Fig. 8, the cause of the operation deviation can be seen.

Specifically, looking at adjustable variable to the direction of PC 2 and investigating the reason that causes operation deviation, it is founded that this is due to the rise in pressure at the HIC1309, LIC1304, HIC1308, PIC1317.

5. PERFORMANCE EVALUATION

Using Excel, exiting function to analyze the process as follows

- Existing Function Model DAO Yield=
  \[(\frac{\text{DA1SDFI1309}}{\text{DA1SDFI1301} + \text{DA1SDFI1302}})\times\text{Function}\]
  \[(\frac{\text{SDA1SDFI1301}}{\text{SDA1SDT1303}}), \frac{\text{SDA1SDFI1307}}{\text{SDA1SDFI1301} - \text{SDA1SDPI1302}}, \text{VDU2A1107}, \text{SDA1SDPI1302}\]

Using existing function model, navigate the variables explanatory power, perform stepwise-multi-
ple regression analysis to eliminate the variables that are not meaningful.

As a result of the analysis, it has been proved that existing function has low explanatory and predictive power. Table 3 shows the result.

Comparing with the above, if multiple regression analysis was performed after changing dependent variable to FIC1309, all the variables affect in a meaningful way. Table 4 shows the result.

Independent variable which has high explanatory power are in the order of AR FEED, VR FEED, TOTAL SOLVENT FLOW, ASPHALT-ENE SEPARATOR TEMPERATURE, ASPHALT-ENE SEPARATOR PRESSURE, AR API.

Explanatory power about FIC1309 is 91% which is 53% of improvement compared to existing model. Table 5 shows the each regression equation and explanatory power based on stage input method.

As the result of this analysis, it has been shown the analysis using ordinary variables can draw improved model. As the existing function has very low predictive power, it can be concluded that the new model offers the better result.

6. CONCLUSIONS

In this paper, MSPC method is suggested to improve operation deviation by using big data which is generated from deasphalting process and the re-

<table>
<thead>
<tr>
<th>Table 3. Result of the existing function analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>FICI</td>
</tr>
<tr>
<td>AR FEED</td>
</tr>
<tr>
<td>VR FEED</td>
</tr>
<tr>
<td>TOTAL SOLVENT FLOW</td>
</tr>
<tr>
<td>ASPHALT-ENE SEPARATOR TEMPERATURE</td>
</tr>
<tr>
<td>ASPHALT-ENE SEPARATOR PRESSURE</td>
</tr>
</tbody>
</table>

All the variables affect in a meaningful way

<table>
<thead>
<tr>
<th>Table 4. The result after changing the SDA to the FIC1309</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>FICI</td>
</tr>
<tr>
<td>AR FEED</td>
</tr>
<tr>
<td>VR FEED</td>
</tr>
<tr>
<td>TOTAL SOLVENT FLOW</td>
</tr>
<tr>
<td>ASPHALT-ENE SEPARATOR TEMPERATURE</td>
</tr>
<tr>
<td>ASPHALT-ENE SEPARATOR PRESSURE</td>
</tr>
</tbody>
</table>

All variables have a significant impact on the FIC1309.

<table>
<thead>
<tr>
<th>Table 5. The result at the SDA process using Output variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>Case 1</td>
</tr>
<tr>
<td>Case 2</td>
</tr>
<tr>
<td>Case 3</td>
</tr>
</tbody>
</table>

The explanatory for the FIC1309 is 91%. On average, the error of the measured value was determined to 2.6%. |
lated application method is also demonstrated.

T-Square chart was used to analyze operation deviation through correlation and exploratory data analysis of dependent and independent variables which related to the applicable process. The result of analysis showed that Operation deviation occurred due to the increase in the pressure of independent variables such as HIC1309, LIC1304.

Big data analysis system that is designed by applying the suggested analysis method can analyze the cause of defective in real time which means it can draw optimized working condition through prediction system. Also optimized process condition can minimize the defect rate, therefore production cost is reduced.

Also by introducing the system for process analysis, analysis time will be shorten which consequentially reduce the cost and the time to analyze the millions of data.

Hence force, if there is continuous research, production efficiency can be maximized in overall oil processing process. Real-time processing analysis will also become feasible by using statistical method for the industry like petrochemical and iron manufacturing business which has streaming process.

REFERENCES


Joo-Hwang Park
He received his B.S. degree from University of Ulsan in 1991, his major is Computer Science. M.S. degree from the department of Software Engineering, Dong-eui University in 2011. He has worked at Enterprise Partner Group in Microsoft Korea as a Manufacturing Team Leader. His current interests are MES and Big Data in manufacturing industry.

Jong-Soo Kim
He received his B.S. degree from Pukyong National University in 1992, his M.S. degree from the department of Computer Engineering, Busan University of Foreign Studies in 2003, and his Ph.D. degree from the department of Software Engineering, Dong-eui University in 2006. Since 2014, he has been a member of the Korea Left College, where he is now the professor in the department of lift engineering. His current research interests are software design and web applications.

Tai-Suk Kim
He received his B.S. degree from the department of electrical engineering, Kyungpook National University in 1981 and his M.S. and Ph.D. degrees from the department of computer science, Keio University in 1989 and 1993, respectively. Since 1994, he has been a faculty member of the Dong-eui University, where he is now the professor in the department of software engineering. His current research interests are information system and Internet business.