CHEMICAL ENGINEERING TRANSACTIONS
VOL. 39, 2014
A publication of AIDIC
The Italian Association of Chemical Engineering
www.aidic.it/cet
Guest Editors: Petar Sabev Varbanov, Jiří Jaromír Klemeš, Peng Yen Liew, Jun Yow Yong

ISBN 978-88-95608-30-3; **ISSN** 2283-9216
DOI:10.3303/CET1439119

Feature Selection Based Root Cause Analysis for Energy Monitoring and Targeting

Tibor Kulcsar[a], Miklos Balaton[b], Laszlo Nagy[b], Janos Abonyi*[a]

[a]University of Pannonia, Department of Process Enginnerng, Egyetem Street 10, H-8200 Veszprem, Hungary;
[b]Hungarian Oil and Gas Company Szazhalombatta, Hungary
janos@abonyilab.com

Energy Monitoring (EM) systems are based on monitoring the difference between targeted and measured energy consumption. Data-driven dynamic targeting models can be used to estimate values of key energy indicators (KEI). In some cases it is difficult to determine which process variables influence the KEIs. We developed an automated root cause analysis (RCA) technique to find the most important driving factors of energy efficiency. The proposed concept is based on the application of feature selection algorithms. We applied Orthogonal Least Squares (OLS) and Random Forest Regression (RFR) to find the proper set of input variables of the targeting models. The concept of the resulted energy monitoring system is applied at the Duna Refinery of MOL Hungarian Oil and Gas Company.

1. Introduction

Advanced production management systems are designed to maximize the production and at the same time minimize cost and emission. Energy portfolio management allows the classification and prioritization of energy consumption to define target-oriented action plans towards energy efficiency improvement (Thiede et al., 2012) A systematic overview of the state of the art in energy and resource efficiency increasing methods and techniques in manufacturing is given in (Duflou, et al., 2012). Energy efficiency has the following four components: performance efficiency, operation efficiency, equipment efficiency, and technology efficiency (Xia and Zhang, 2010). In our paper we focus on the improvement of operation efficiency.

Energy monitoring improves operational energy efficiency by continuous comparison of actual and estimated energy consumption. Methods for calculating expected consumption fall into two categories. Precedent based methods make comparisons of actual energy consumption with previous periods (Behrendt et al., 2012), while activity-based methods calculate expected values of key energy indicators (KEI) from the relevant process variables (Abonyi and Kulcsar, 2013). Understanding the effects of these driving factors has significant economic and technological potential, e.g. such knowledge is also valuable to support Process Integration (Chew et al., 2013).

In some cases it is difficult to determine which process variables influence the KEIs. In these situations the input variables of the targeting models should be selected based on root cause analysis of the operation of the technology. Unfortunately this procedure is subjective and time-consuming and does not guarantee a model with good prediction performance.

Root Cause Analysis (RCA) is a method of problem solving that tries to identify the root causes of faults and problems. We applied the RCA approach to find the driving factors of energy efficiency of process plants. There are many ways to implement RCA. For example Bayesian networks can be applied to find the root causes of deviations during the operation of complex processes (Weidl et al., 2005). Digraph models were proven to be useful to identify discrete events (faults) (Wan et al., 2013). Multivariate statistical process monitoring (MSPM) with some extensions is a useful technique to isolate not only the effects of the faults, but also the underlying causes. For this purpose MSPM and fuzzy-signed directed graphs were combined to identify the root causes (Ha et al., 2014). These methods have in common that each is developed for discrete event systems.

Please cite this article as: Kulcsar T., Balaton M., Nagy L., Abonyi J., 2014, Feature selection based root cause analysis for energy monitoring and targeting, Chemical Engineering Transactions, 39, 709-714 DOI:10.3303/CET1439119

Building energy monitoring models requires the knowledge of the the driving factors of the energy efficiency (Abonyi and Kulcsar, 2013). The above mentioned techniques are designed to analyse discrete events not for handling continuous process variables. To support root cause analysis of energy efficiency we proposed a fully automated feature selection based approach. The methodology is based on the application of Orthogonal Least Squares (OLS) and Random Forest Regression (RFR) to find the proper set of input variables of the targeting models from the historical data of hundreds process variables.
The concept of the resulted energy monitoring system is applied at the AV2 unit of the Danube Refinery of MOL Hungarian Oil and Gas Company. The Key Energy Indicators were calculated based on one-year historical data as we assumed that the range of this dataset is wide enough to cover operation ranges of high and low energy consumptions and contains information about the significant malfunctions. The results show that the proposed approach is able to determine useful and informative sets of driving factors of the energy efficiency.

2. Targeting model based energy monitoring

Activity-based energy targets are usually calculated by linear regression models,

$$\widehat{y_k} = \boldsymbol{x_k^T \theta} \tag{1}$$

where the calculated output $\hat{y}_i$ is the linear combination of process variables (drivers), $\boldsymbol{x}_k = [x_{1,k}, \dots, x_{n,k}]$, where *k* represents the k-th sampling time and *n* stands for the number of process variables having significant effect to the energy consumption. At the development of this model it is important to ensure that data are synchronised as closely as possible with the required assessment intervals. Based on a synchronized set of data $\{y_k, \boldsymbol{x}_k\}, k = 1, \dots, N$ linear least squares method can be applied to find optimal parameters of the model $\boldsymbol{\theta}$ that minimizes the $\sum(y_k - \widehat{y_k})^2$ quadratic cost function.

$$\boldsymbol{\theta} = (\mathbf{X^T X})^{-1}\mathbf{X}^T\mathbf{y} \tag{2}$$

where $\mathbf{X}$ is $n \times N$ matrix of historical process variables and $\mathbf{y}$ is an $N \times 1$ vector of measured output variable (energy consumption or efficiency measure).When the predicted consumption $\widehat{y_k}$ is higher as the measured value y_i the technology is considered to be efficient regards to historical data. The relation $\widehat{y_k} < y_k$ suggests that the technology could work with lower energy consumption.

3. Orthogonal Least Squares based Feature Selection

The performance of data-driven targeting models dependents on complex set of process variables. When no proper prior knowledge is available for the selection of the driving factors of a KEI model, feature selection algorithms can be used for sophisticated and automated root cause analysis.
The OLS algorithm is an effective tool to determine which terms are significant in a linear-in-parameters model, since it is based on the error reduction ratio (err) which is a measure of the decrease in the variance of output by a given term. In the following the details of this algorithm are presented.
The compact matrix form corresponding to the linear-in-parameters model (1) is $\boldsymbol{y} = \boldsymbol{X\theta} + \boldsymbol{e}$, where the $\boldsymbol{X}$ is the regression matrix (2), $\boldsymbol{\theta}$ is the parameter vector, $\boldsymbol{e}$ is the error vector. The OLS technique transforms the columns of the $\boldsymbol{X}$ matrix (2) into a set of orthogonal basis vectors in order to inspect the individual contributions of each term.
The OLS algorithm assumes that the regression matrix $\boldsymbol{X}$ can be orthogonally decomposed as $\boldsymbol{X} = \boldsymbol{WA}$, where $\boldsymbol{A}$ is an $n \times n$ upper triangular matrix (it means $A_{i,j} = 0$ if $i > j$) and $\boldsymbol{W}$ is an $N \times n$ matrix with orthogonal columns in the sense that $\boldsymbol{W^T W} = \boldsymbol{D}$ is a diagonal matrix. (N is the length of $\boldsymbol{y}$ vector and n is the number of regressors.) After this decomposition one can calculate the OLS auxiliary parameter vector $\boldsymbol{g}$ as

$$\boldsymbol{g} = \boldsymbol{D}^{-1}\boldsymbol{W}^T\boldsymbol{y} \tag{3}$$

where g_i is the corresponding element of the OLS solution vector.
The output variance $(\boldsymbol{y}^T\boldsymbol{y})$ can be explained as

$$\boldsymbol{y}^T\boldsymbol{y} = \sum_{i=1}^{M} g_i^2 w_i^T w_i + \boldsymbol{e}^T\boldsymbol{e} \tag{4}$$

Thus the error reduction ratio, $[err]_i$ of the i-th input variable can be expressed as $[err]_i = \frac{g_i^2 w_i^T w_i}{\mathbf{y}^T \mathbf{y}}$.

This ratio offers a simple mean to order and select the model terms of a linear-in-parameters model according to their contribution to the performance of the model.

4. Random Forest Regression Based Feature Selection

The drawback of OLS is that it assumes linear relationship between inputs and the output. Regression trees are simple, transparent and easily interpretable nonlinear models. The combination of these trees results a forest of the models. When the regression trees are statistically independent, the average of the prediction of these models will be better than the prediction of the individual models. Furthermore, the analysis of the forest can be used to select the most important process variables. In the following the theoretical background of this technique will be presented.

The concept of random forest was developed by Leo Breiman (Breiman, 2001). Andy Liaw implemented Breiman's concept in R (Liaw, 2012). We used the MATLAB hosted version of this R package. The method combines Breiman's "bagging" idea and the random selection of features. Random forests for regression are formed by growing trees depending on random matrix $\boldsymbol{\Theta}$**.** The $\boldsymbol{\Theta}$ consist of a number of independent random integers between 1 and M, where M is the number of trees in the forest. The nature and dimensionality of $\boldsymbol{\Theta}$ depends on its use in the tree construction.

A random forest is a predictor consisting a collection of tree-structured predictors $\{h_i(\boldsymbol{x}, \boldsymbol{\Theta}_i), i = 1, \dots\}$ where the $\{\Theta_k\}$ are independent identically distributed random vectors and each tree cast a unique estimation for output $\hat{y}$ at input $\boldsymbol{x}$**.** The output values are numerical and we assume that the training set is independently drawn from the distribution of the given $\boldsymbol{y}, \boldsymbol{X}$ dataset. The mean-squared generalization error for any numerical predictor $h_i(\boldsymbol{X}) = h(\boldsymbol{X}, \boldsymbol{\Theta}_i)$ is

$$E_{X,y}\left(\left(y - h_i(X)\right)^2\right). \tag{5}$$

where $E_{\boldsymbol{X},\boldsymbol{y}}$ denotes the expected value and $\left(y - h(X)\right)^2 = \left(y - h(X)\right)^T \left(y - h(X)\right)$, and we use this substitution in the following. The random forest predictor is formed by taking the average over M of the trees $\{h(\boldsymbol{x}, \boldsymbol{\Theta}_i)\}$. Use of the proof (Breiman, 2001) of Almost Sure Convergence theorem, as the number of trees in the forest goes to infinity, mean-squared generalization error goes to a limit value almost surely as:

$$E_{\boldsymbol{X},\boldsymbol{y}}\left(\left(\boldsymbol{y} - \frac{\sum_{i=1}^{M} h(\boldsymbol{X}, \boldsymbol{\Theta}_i)}{M}\right)^2\right) \to E_{\boldsymbol{X},\boldsymbol{y}}\left(\left(\boldsymbol{y} - E_{\boldsymbol{\Theta}}\left(h(\boldsymbol{X}, \boldsymbol{\Theta})\right)\right)^2\right) \tag{6}$$

Denote the right hand side (limit value) of (6) as $PE^*(forest)$ - the generalization error of the forest. Define the average generalization error of a tree as:

$$PE^*(tree) = E_{\Theta}\left(E_{\boldsymbol{X},\boldsymbol{y}}\left(\left(\boldsymbol{y} - h(\boldsymbol{X}, \boldsymbol{\Theta})\right)^2\right)\right) \tag{7}$$

The concept is based on the fact $PE^*(forest) < \bar{\rho} PE^*(tree)$, where $\bar{\rho}$ is the weighted correlation between the residuals $\left(\boldsymbol{y} - h(\boldsymbol{X}, \boldsymbol{\Theta})\right)$ and $\left(\boldsymbol{y} - h(\boldsymbol{X}, \boldsymbol{\Theta}')\right)$, where $\boldsymbol{\Theta}, \boldsymbol{\Theta}'$ are independent, as the weighted correlation is defined as:

$$\bar{\rho} = E_{\boldsymbol{\Theta}}\left(E_{\boldsymbol{\Theta}'}\left(\rho(\boldsymbol{\Theta}, \boldsymbol{\Theta}') S(\boldsymbol{\Theta}) S(\boldsymbol{\Theta}')\right)\right) / \left(E_{\boldsymbol{\Theta}}\left(S(\boldsymbol{\Theta})\right)\right)^2 \tag{8}$$

where $S(\boldsymbol{\Theta}) = \sqrt{E_{\boldsymbol{X},\boldsymbol{y}}\left(\left(\boldsymbol{y} - h(\boldsymbol{X}, \boldsymbol{\Theta})\right)^2\right)}$ is the standard deviance of prediction errors.

To obtain accurate regression forest this theorem requires low correlation between residuals and low error trees. The random forest decreases the average error of the trees employed by the factor $\bar{\rho}$. The randomization employed needs to aim at low correlation. (Breiman, 2001)

To rank the process variables and select a proper subset we used the importance measures which are defined on the following way. The first measure is computed from random permutating the data: For each tree, the prediction error (MSE) is recorded. Then the same is done after permutation each predictor variables. The difference between the two are then averaged over the trees, and normalized by the

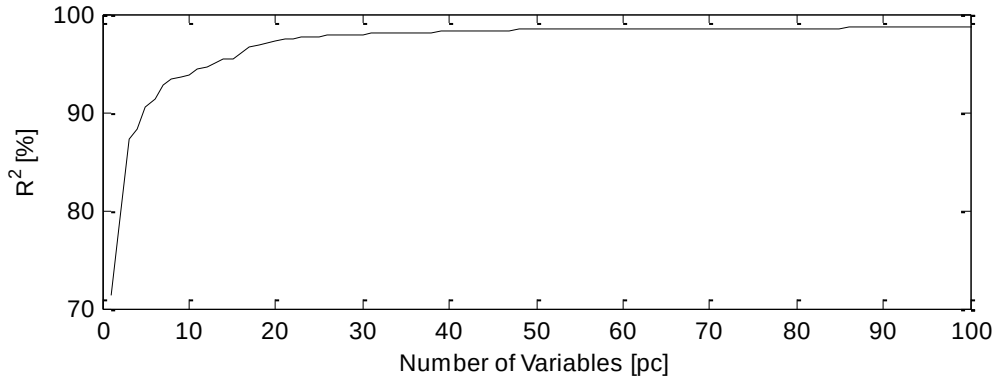

Figure 1: Model accuracy for fuel gas consumption in function of the increasing number of the relevant input variables

standard deviation of differences. If the standard deviation of the differences is equal to 0 for a variable, the division is not done (but the average is almost always equal to 0 in that case). (Breiman et al., 2012). The second measure is the total decrease in node impurities from splitting on a variable, averaged over all trees and it is measured by the residual sum of squares. (Breiman et al., 2012).

5. Results

The proposed technique is applied to support the targeting model development project of the MOL Hungarian Oil and Gas Company. In this paper results related to two Key Energy Indicators (KEIs) of AV2 plant are presented. The applicability of the orthogonal least squares based feature selection is demonstrated on the total fuel gas consumption of the furnaces of the AV2 plant, while the random forest based feature selection is applied to model the plant wide electric power consumption of AV2.

5.1 OLS based Feature Selection on Fuel Gas Consumption

The OLS model was used to find the most relevant variables influencing the gas consumption of the furnaces among 620 historical process variables. Figure 1 shows how the accuracy of the model increases by adding more and more input variables as the variables are introduced to the model by the decreasing series of relevance given by OLS. The model performance is measured by the model correlation (R^2). Figure 1 shows that the model's performance which is built using the first two most relevant variables has already $R^2 = 0.88$ correlation.

The first five variables given by OLS gives a compact yet accurate model, the fuel gas consumption can be predicted with $R^2 = 0.92$. These most important variables are:

1. Main boiler temperature
2. Temperature of heating steam
3. Liquid level in the main boiler
4. Density of fuel gas
5. Total crude oil feed

This list of variables reflects the knowledge and expertise of the process engineers.

However, it should be noted that statistical correlation does not necessarily results in informative features, we often neglected statistically informative variables from the final model based on the suggestions of the engineers. Therefore, the proposed tool should be handled only as tool for decision support. A proper way to use OLS based feature selection is the following:

1. Let OLS to select a large set of variables.
2. Among these potential inputs select a smaller set based on prior knowledge of the process.

5.2 RF based Feature Selection on Electric Power Consumption

We used random forest feature selection to select a proper set of variables which are relevant to the complete electric power consumption (KEI) of AV2 unit. Based on prior knowledge of the process engineers we know that almost all the electric power is consumed by the main process pumps (total feed, inter tower streams, cooling water and product streams). Based on this prior knowledge we expect that the feature selection algorithm should highlight the importance of flow rates and pressures.

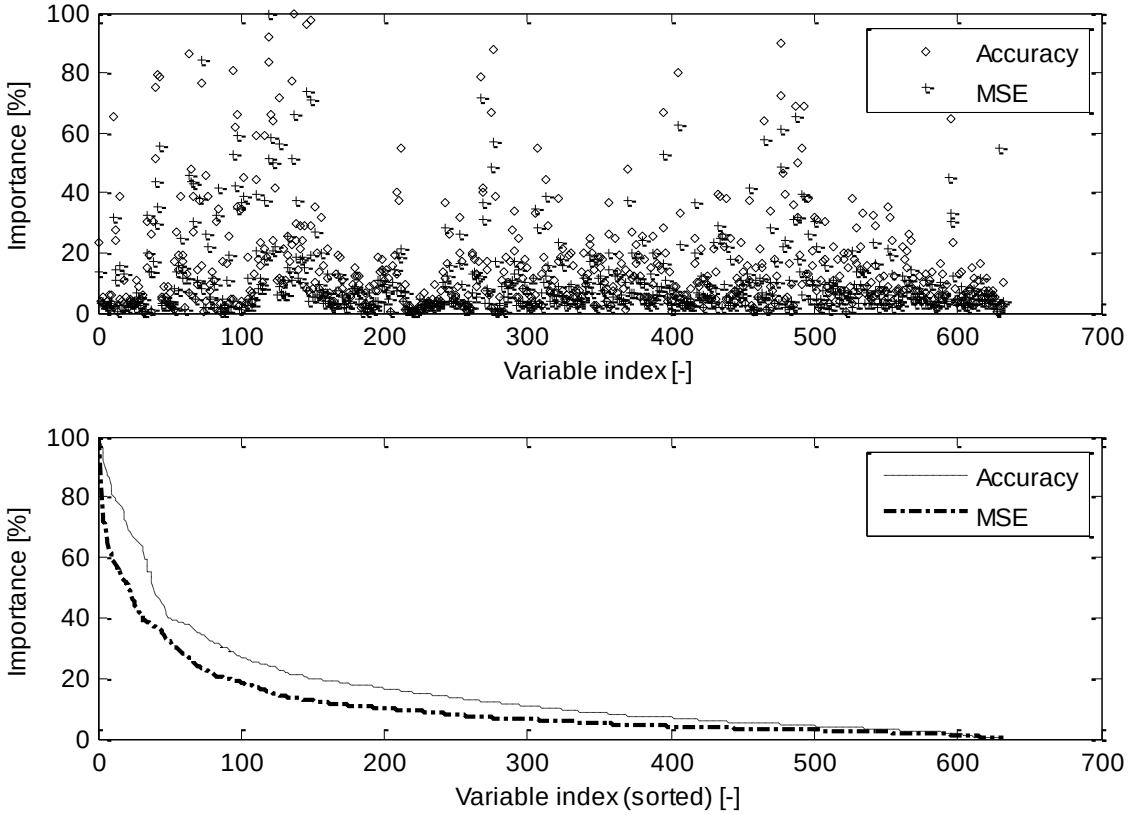

Figure 2: Relevance of process variables given by random forest regression

For our calculations we used the MATLAB hosted R package implementation of the FORTRAN77 program created by Leo Breiman. The forest contained 500 regression trees. Each tree was grown using five randomly selected process variables from the original variable set. Figure 2 shows the normalized importance of each variables in alphabetical order (top), and ordered according to their importance level (bottom). As the results show, the relevance of variables is decreasing exponentially.
The total crude oil feed, the inlet pipe pressures and the flows of main process streams were proven the most important variables, which ordering was also confirmed by the process engineers.
We analyzed the prediction performance of the random forest using validation samples. On the validation set the model correlation was excellent, $R^2 = 0.97$.
The selected variables were also used to formulate a linear model. The linear model with the ten most significant variables was also quite accurate, $R^2 = 0.92$, as this accuracy is better that suggested in the patent related to feature selection for energy monitoring (Resina, 2006).

6. Conclusions

Energy Monitoring is based on monitoring the difference between targeted and measured energy consumption. In some cases it is problematic develop accurate and informative targeting models, since it is difficult to determine which process variables influence the KEIs. We developed an automated root cause analysis (RCA) technique to find the most important driving factors of energy efficiency. The proposed concept is based on the application of feature selection algorithms. We examined two regression methods with feature selection capability for energy monitoring applications. We applied orthogonal least squares regression and random forest regression to predict key energy indicators and select the most important process variables which are relevant to the KEIs. The applicability of these methods was demonstrated on two KEI of AV2 plant in MOL Duna Refinery. Based on the results we can conclude that both methods are able to predict the KEI values and select the most relevant process variables.

Acknowledgement

This research of Janos Abonyi was supported by the European Union and the State of Hungary, co-financed by the European Social Fund in the framework of TÁMOP 4.2.4.A/2-11-1-2012-0001 'National Excellence Program'. The infrastructure of the research was supported by the TAMOP-4.2.2/A-11/1/KONV-2012-0071 project.

References

Duflou J.R., Sutherland J.W., Dornfeld D., Herrmann C., Jeswiet J., Kara S., Hauschild M., Kellens K., 2012. Towards energy and resource efficient manufacturing: A processes and systems approach. CIRP Annals - Manufacturing Technology, 61(2), 587–609. DOI:10.1016/j.cirp.2012.05.002.

Thiede S., Bogdanski G., Herrmann C., 2012. A Systematic Method for Increasing the Energy and Resource Efficiency in Manufacturing Companies. Procedia CIRP, 2, 28–33. DOI:10.1016/j.procir.2012.05.034.

Xia X., Zhang J., 2010. Energy efficiency and control systems-from a POET perspective. Methodologies and Technology for Energy Efficiency. <www.ee.up.ac.za/main/_media/en/ postgrad/subjects/ees732/ 2010portugalpaper124.pdf> accessed 25/05/2014.

Chew K.H., Alwi S.R.W., Klemeš J.J., Manan Z.A.2013. Process modification potentials for total site heat integration. Chemical Engineering Transactions, 35, 175-180.

Weidl G., Madsen A.L., Israelson S., 2005. Applications of object-oriented Bayesian networks for condition monitoring, root cause analysis and decision support on operation of complex continuous processes, Computers and Chemical Engineering 29, 1996-2009

Wan Y., Yang F., Lv N., Xu H., Ye H., Li W., Xu P., Song L., Usadi A.K., 2013. Statistical root cause analysis of novel faults based on digraph models, Chemical Engineering Research and Design 91, 87-99

He B., Chen T., Yang X., 2014, Root cause analysis in multivariate statistical process monitoring: Integrating reconstruction-based multivariate contribution analysis with fuzzy-signed directed graphs, Computers and Chemical Engineering 64, 167–177

Abonyi J., Kulcsar T., Balaton M., Nagy L., 2013. Historical Process Data Based Energy Monitoring -Model Based Time-Series Segmentation to Determine Target Values., Chemical Engineering Transactions, 35, 931-936, DOI: 10.3303/CET1335155.

Breiman L., 2001. Random Forests. Machine Learning, 45 (1), 5–32. DOI:10.1023/A:1010933404324.

Liaw A., 2012. Documentation for R package randomForest, <cran.r-project.org/web/packages/randomForest/randomForest.pdf> accessed on 29/08/2013

Retsina T. 2006, Method and System for Targeting and Monitoring the Energy Performance of Manufacturing Facilities. US patent US 7,103,452 B2.

Behrendt T., Zein A., Min S., 2012. Development of an energy consumption monitoring procedure for machine tools. CIRP Annals - Manufacturing Technology, 61(1), 43–46. DOI:10.1016/j.cirp.2012.03.103.

Breiman L., Cutler A., Liaw A., Wiener M., 2012. Random Forest Package – Breiman and Cutler's random forest for classification and regression.,<stat-www.berkeley.edu/users/breiman/RandomForests/>, accessed 25/05/2014.