CHEMICAL ENGINEERING TRANSACTIONS

VOL. 36, 2014

A publication of
AIDIC
The Italian Association
of Chemical Engineering
www.aidic.it/cet

Guest Editors: Valerio Cozzani, Eddy de Rademaeker
Copyright © 2014, AIDIC Servizi S.r.l.,
ISBN 978-88-95608-27-3; **ISSN** 2283-9216

DOI: 10.3303/CET1436035

Method of Selecting Process Signals for Creating Diagnostic Machines Optimised to Detect Abnormalities in a Plant Using a Support Vector Machine

Hirotsugu Minowa*[a], Yoshiomi Munesawa[b], Yuichiro Furuta[c], Akio Gofuku[a]

[a] Graduate School of Natural Science and Technology, Okayama University, Okayama, Japan
[b] Engineering Department, Hiroshima Institute of Technology, Hiroshima, Japan
[c] NTT Comware Corporation, Tokyo, Japan
minowa-h@okayama-u.ac.jp

In industrial plants, a multi-agent diagnostic system using machine learning with high discrimination performance to prevent the recurrence of accidents are desired because the damages to the company, industrial world, and humans are serious if an accident occurs. However, machine learning needs to be optimised according to the target, e.g. a plant operating condition diagnosis. Thus, we propose a method that selects process signals to optimise the performance of a diagnostic machine for a plant using factor analysis and decision-tree analysis. A feature of our optimisation method is that it accounts for combinations of process signals. Further, an advantage of our method is that the time to create an optimised diagnostic machine is short. The diagnostic machines need to be updated in a limited period, e.g. plant equipment repair, which changes regular process values. Another advantage is that our method can be applied to various learning machines to improve performance. This advantage allows the designer of the diagnostic system to use the best machine-learning method on each diagnostic machine. This paper reports our methodology, our proposed method, and the experimental results where a diagnostic machine was improved by 5.3% to 98.8% from 93.5% abnormality detection accuracy when our method which implemented a support vector machine was applied to the diagnostic machine to detect the blockage of a pipe in the desulfurisation process in a chemical plant simulator.

1. Introduction

Studies using machine learning to diagnose an operating condition by analysing a process signal (below, signal) in an industrial plant have been widely performed to prevent the recurrence of accidents (Widodo, 2007, Ardi, 2009, Munesawa, 2013). Machine-learning methods are generally superior to conventional methods such as principal component analysis (PCA) or methods of detection by thresholds. Therefore, researchers have studied machine learning to diagnose plant operating conditions more effectively. There are two types of machine-learning methods. One is supervised machine learning, and the other is unsupervised machine learning. Unless stated otherwise, ‘machine learning’ in this paper’ refers to ‘supervised machine learning’ such as a support vector machine (SVM) (Vapnik, 1995).

However, there is a problem in improving the performance of machine learning. There are two types of abilities in machine learning. One is the classification ability to detect a known abnormal condition (below, abnormal condition); the other is the generalisation ability to detect an unknown abnormal condition. The problem is that the improvement of these two abilities is exclusive.

To solve this problem, a multi-agent diagnostic system (MADS), e.g. hybrid fault diagnosis, is used (Akio, 2013). One conventional diagnostic machine is responsible for the two roles of generalisation and classification. However, improvement of both the classification and generalisation performance is difficult owing to the exclusive performance limitations. The advantage of an MADS is that it can separate the two roles of generalisation and classification in a one diagnostic machine because an MADS can have two groups with diagnostic machines which are responsible for either classification or generalisation. This advantage overcomes the exclusive limitation of machine learning because it allows diagnostic machine

Please cite this article as: Minowa H., Munesawa Y., Furuta Y., Gofuku A., 2014, Method of selecting process signals for creating diagnostic machine optimized to detect an abnormalities in a plant using support vector machine, Chemical Engineering Transactions, 36, 205-210 DOI: 10.3303/CET1436035

specialisation to improve the performance of either classification or generalisation. In addition, a framework can install the diagnostic machines created by various types of machine learning, allowing the designer of the diagnostic system to select the best machine-learning techniques for diagnosis.

There are some methods to improve the diagnostic performance of MADSs, such as the method proposed by Hiroyasu (2008) which improves the both classification and generalisation, but the MADS allows the diagnostic machine to improve only either classification or generalisation by the separation of roles. The method proposed by You (2009) changes the internal mechanism; however, this method cannot be applied to other machine-learning techniques. In addition, there is no method that can optimise the classification performance on the basis of multiple combinations of signals. The optimisation needs to be maximised as much as possible within the limited maintenance time. Furthermore, optimisation is desirable for the applications of various types of machine learning.

Therefore, we propose a method to optimise the performance of a diagnostic system for an MADS. Our method aims at optimising the classification performance considering signal combinations to complete the optimisation within a limited time as soon as possible. In addition, our method has features which can be applied to diagnostic machines of various types of machine learning. We report our methodology, the proposed signal selection method, and the results of the evaluation experiments.

2. Our methodology

A machine-learning diagnostic machine is a calculator which can classify an operating condition (Normal / Abnormal) according to the input signal value. The machine learning has a feature when the value ranges of the signal are more clearly separated according to the operating condition, a greater improvement in the classification performance of a diagnostic machine which created from these signals is realised. The performance of this diagnostic machine can be expected to be higher in case that generating a diagnostic machine by using signals which range are separated in accordance with the operating conditions. Therefore, our method uses factor analysis to find the signals which ranges are separated by operating condition. The factor analysis is a method to measure the factor loadings as correlation values at each factor, which are useful to find a group of common factor. A higher contribution rate for a factor indicates that signals included in it are more similar. The number of signals which the displacements were changed according to the time of an abnormality occurrence ought be most because the change of the process value propagates to the other signals. Therefore, our optimizing method selects the signals of the highest contribution rates which has a lot of signals which values changes according to the timing of an accident occurrances to make the best diagnostic machine.

In addition, Our method selects signals which selected by factor analysis according to the performance of the diagnostic machines created from combinations of multiple their signals. The diagnostic performance of the diagnostic machine is determined by a multi-dimensional mapping space (MDMS) made from beneficial combinations of signals. The performance of the diagnostic machine declines if the MDMS of it is generated by the signals involves noise. The way to measure the benefit of the combination of signals for diagnosis have no choice but to measure the performance of the diagnostic machine made from the combination of signals every time its combination changed. Therefore, the time which to obtain the most beneficial combination of signals by measuring performance of all diagnostic machine is increases exponentially in proportion with the number of signal combinations.

Thus, our method measures the diagnostic perfomance of the combinations of a few signals and finds the beneficial signal combination by decision-tree analysis. The decision-tree analysis is a method of grouping targets by the values of each attribute of the target by creating a decision tree. Our method uses decision-tree analysis to find signal groups with a high model score according to relationship between the model scores and a few signals.

2.1 Model score

This section explains the model score, which is the diagnostic performance score of a diagnostic machine. The model score is an average of the percentage of correct answers in the diagnosis of normal and abnormal operating conditions. The model score is calculated by Eq(1).

$$M = \frac{1}{w_n + w_a}\left(w_n \frac{n_{An}}{n_{Sn}} + w_a \frac{n_{Aa}}{n_{Sa}}\right) \tag{1}$$

In Eq(1), n_{Sn} and n_{Sa} are the total numbers of recorded normal and abnormal conditions, respectively; n_{An} and n_{Aa} are the recorded numbers of answers for the diagnosis of normal and abnormal conditions which were correct, respectively; and w_n and w_a are the weight coefficients. In this study, $w_n = w_a$. If a diagnostic machine which detects abnormalities more correctly is desired, w_a should be large.

3. Method

The proposed method consists of two steps. Before explaining the algorithm of our method, we define the following variables. n process signals are defined as $S_i (1 \leq i \leq n)$. The values of each k $(1 \leq k \leq K)$ time step in S_i are defined as x_{ik}, the plant operating conditions of each time step k are defined as y_k, and the data set D_T is defined as a 2D vector $(x_{1k}, \dots, x_{nk}, y_k)$ consisting of S_i as an explanatory variable and y_i as an objective variable. A function $M(D_T)$ returns a model score which is calculated by evaluating each record of each time stamp in D_T by the diagnostic machine created from D_T.

3.1 Step 1

Step 1 of our method selects beneficial signals to find signals of which values change following an abnormality occurence by factor analysis and decision-tree analysis.

Step 1-1 Normalisation

The value range of each signal is different. This normalisation process changes the value range of signals to one where an average of zero and a variance is applied to D_T to obtain the correct results from the factor analysis. The normalisation is defined as Eq(2).

$$z_{ik} = \frac{x_{ik} - \overline{x_\iota}}{sd(x_i)} \tag{2}$$

In Eq(2), $sd(x_i)$ is the standard deviation, and $\overline{x_\iota}$ is the average of the value x_{ik} in each signal S_i. $D_T N$ is the data set in which D_T was normalised, which consists of the vector $(z_{1k}, \dots, z_{nk}, y_k)$.

Step 1-2 Applying the factor analysis

The factor analysis needs to set the number m of factors F as an analysing parameter. The contribution rates are defined as $c_j (1 \leq j \leq m, c_1 \geq c_2 \geq \cdots \geq c_m)$ of each factor F_j. m increases one by one gradually and is determined when $\sum_{j=1}^{m} c_j$ is larger than the threshold Th. The factor analysis of the varimax rotation method is applied to $D_T N$ which calculates the factor loadings f_{ji} of each signal S_i of each factor F_j.

Step 1-3 Grouping the process signals

A decision-tree analysis is applied to the absolute values of the factor loadings f_1 on the factor F_1 which has the highest contribution rate c_1. The signals classified into a leaf nodes, which are the end point of the decision tree, are defined as $SG_h (1 \leq h \leq n_{G1})$. Here, n_{G1} is defined as the number of the groups.

Step 1-4 Calculation of the base model score

The base model score M_{B1} is $M(D_T N)$.

Step 1-5 Calculation of the model scores of each leaf node.

The model scores $M_h (1 \leq h \leq n_{G1})$ are $M(SG_h)$.

Step 1-6 Exclusion of invalid process signals

Signals of the groups SG_h of which the value M_h are smaller than M_{B1} are removed. The remaining signals are defined as D_{T2} and analysed in Step 2. The number of remaining signals in D_{T2} is defined as n_2.

3.2 Step 2

Step 2 of our method selects the beneficial signals by measuring strictly the model score of the diagnostic machines created from a combination of signals.

Step 2-1 Calculation of the base model score

The base model score M_{B2} is calculated from $M(D_{T2})$.

Step 2-2 Evaluation of the diagnostic performance from the combination of two signals

Signal combinations $Com(D_{T2}, 2)_k (1 \leq \mathrm{o} \leq q = {}_{n_2}C_2)$ are combinations of two signals selected from D_{T2}. The data set D_{T3} is defined as the vector $(Com(D_{T2}, 2)_\mathrm{o}, y_\mathrm{o})$.The model scores M_o^2 are calculated from $M(D_{T3})$. D_{T4} is defined as the vector {$(Com(D_{T2}, 2)_1, \mathrm{M}_1^2)$, $(Com(D_{T2}, 2)_2, \mathrm{M}_2^2)$,…, $(Com(D_{T2}, 2)_q, \mathrm{M}_q^2)$}.

Step 2-3 Application of the decision-tree analysis

A decision-tree analysis is applied to D_{T4}, and the decision tree is obtained. The number of the leaf node is defined as n_3.The leaf nodes of the decision tree are defined as $LS_l(1 \leq l \leq n_3)$. The model scores of the signals of each leaf node LS_l are defined as M_l^4.

Step 2-4 Selection of the beneficial signals
The selected signals S_o^2 of which $M_{n_3}^3$ are greater than M_{B2} are selected as in Eq(3).

$$\{S_o^2 | \, M_{n_3}^3 > M_{B2}\} \tag{3}$$

4. Experimental evaluation

An experimental evaluation was carried out to measure the performance of the optimisation and the effectiveness of our method applied to an assumed accident in an industrial plant.

4.1 Experimental environment

The decision-tree analysis used a method based on the CART algorithm. The SVM method was C-classification, and the kernel of the SVM was the RBF algorithm. An script for the evaluation experiment was created by the R language ver. 2.15.3 in Windows 7. The decision-tree analyses were executed by the mvpart plug-in of the R language. The SVM methods were executed by the e1071 plug-in.

4.2 Process details

The experimental evaluation was performed for the desulfurisation process shown in Figure 1. The desulfurisation process is a process that separates the input material LPG into propane and butane in a petrochemical plant. The material is separated by a vapor–liquid equilibrium operation as steam boils the material in a reboiler located in the bottom of the tower. The light distillate separated from the material is extracted from the top of the tower by cooling using a heat exchanger. A part of the light distillate returns to the distillation column in reflux. The heavy fraction is discharged from the bottom of the tower. The categories of the sensors in Figure 1 are described in Table 1.

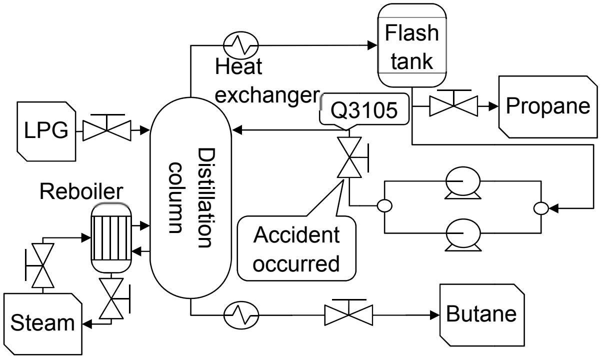

Figure 1: Process of desulfurisation

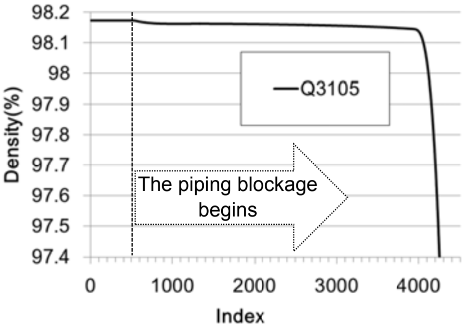

Figure 2: Profile of Q3105

Table 1: Process signals in desulfurisation process

Type of process signal	Name of process signal
Flow sensor	FI3101, FI3102, FI3103, FI3104, FI3105
Liquid level sensor	LI3101, LI3102, LI3103
Concentration sensor	i105, i106, QI3105
Pressure sensor	PI3101, PI3102, PI3103, PI3104, PI3105
Temperature sensor	TI3101, TI3102, TI3103, TI3104, TI3105, TI3106

4.3 Assumed abnormal conditions

It is assumed that the adhesion of impurities causes a piping blockage that inhibits the propane flow from the flash tank to the distillation column. This accident supposes to occur near a concentration meter Q3105 located at near the safety valve on the propane flow. The meter shows the concentration of propane flow, which is an inflow into the distillation column. Figure 2 shows the profile of the Q3105 concentration meter. In Figure 2 the process value begins at about 98.2% as steady-state. The piping blockage as abnormality begins at the 502nd index where a vertical line drawn in Figure 2. The 502nd index begins the gradual decrease of the material and the 4000nd begins the rapid decrease of the material suddenly. We aims for our proposed system to detect the anomaly before the value of concentration in Q3105 is less than 98%, at which the distillate can no longer be shipped as product. The process signals were acquired from Visual Modeler ver. 2.4 as a chemical plant simulator. The sampling rate of the simulation was about 1–2 s.

4.4 Applying our method

Step 1-1 was applied to normalise D_T and obtains D_TN. A factor analysis with $m = 4$ determined when $Th \geq 90$ was applied to $\mathrm{D_T}$ according to Step 1-2. The factor analysis contribution rates were $\mathrm{C_1}$ =11.2% at F_1, $\mathrm{C_2}$ =6.3% at F_2, $\mathrm{C_3}$ =4.5% at F_3, and $\mathrm{C_4}$ =4% at F_4. According to Step 1-3, a decision-tree analysis was applied to the absolute value of the factor loading f_1 which has the highest c_1. These results created two signal groups. Signal group SG_1 consists of seven signals: FI3101, LI3101, LI3102, LI3103, PI3101, PI3102, and TI3106, which are shown in Figure 3. SG_2 consists of 15 signals: FI3102, FI3103, FI3104, FI3105, i105, i106, QI3105, PI3103, PI3104, PI3105, TI3101, TI3102, TI3103, TI3104, and TI3105, which are shown in Figure 4. Step 1-4 calculated the base model score M_{B1} as 0.935 ($w_n \frac{n_{An}}{n_{Sn}} = 1.0$, $w_a \frac{n_{Aa}}{n_{Sa}} = 0.867$), as calculated from $M(D_TN)$. According to Step 1-5, M_1 of the evaluated signals in SG_1 was 0.5 ($w_n \frac{n_{An}}{n_{Sn}} = 0$, $w_a \frac{n_{Aa}}{n_{Sa}} = 1$), and M_2 of the evaluated signals in SG_2 was 0.952 ($w_n \frac{n_{An}}{n_{Sn}} = 1$, $w_a \frac{n_{Aa}}{n_{Sa}} = 0.903$). According to Step 1-6, the signals in SG_1 were removed because the model score M_1 = 0.5 was smaller than M_{B1}. The signals in SG_2 remained for Step 2, and D_{T2} was defined as a 2D vector ($\mathrm{S_i} \in \mathrm{SG_2}, y$).

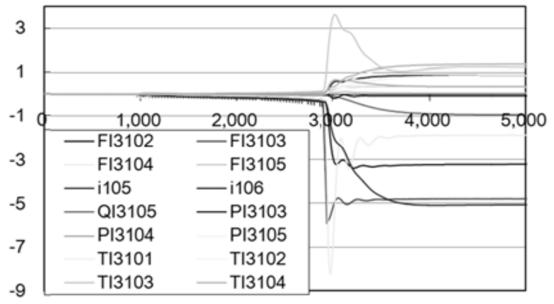

Figure 3: Profile of signals of the group SG1

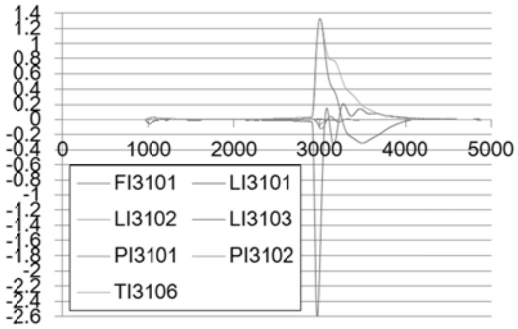

Figure 4: Profile of signals of the group SG2

Step 2-1 calculated a model score $\mathrm{M_{B2}} = 0.952$, which is $\mathrm{M(D_{T2})}$. According to Step 2-2, combinations of two signals $Com(D_{T2}, 2)$ totalling $105(= {}_{15}C_2)$ were created. D_{T3} was consist of $(Com(D_{T2}, 2), y_k)$. The model score M_{h2} was calculated from D_{T3}. Step 2-3 makes a decisison-tree analysis applied to D_{T4}. Figure 5 shows the results of the decision-tree analysis. The leaf nodes LS_l , $n_3 = 6$, consist of nodes 6–11 in Table 5. M_l^4 was calculated from each LS_l. Step 2-4 selected the beneficial signals in nodes 6, 7, 10, and 11 in LS_l because the model scores $M_{n_3}^3$ of the signals of each node 6, 7, 10, and 11 were larger than M_{B2}. The 11 signals selected as an answer were: FI3102, FI3103, FI3104, FI3105, i105, i106, PI3103, PI3104, PI3105, TI3104, and TI3105. The model score was 0.998 ($w_n \frac{n_{An}}{n_{Sn}} = 1$, $w_a \frac{n_{Aa}}{n_{Sa}} = 0.996$) on an optimised diagnostic machine. Our method improved the model score by 0.053 from 0.935 to 0.988.

Table 5: Results of decision-tree analysis.

```
N=105, format: num. of node), split, n, deviance, yval * denotes terminal node
1) root 105 3.859349e-02 0.9869138
 ├ 2) Signal1=QI3105, TI3101, TI3102, TI3103 6 1.644991e-03 0.9173634
 │  ├ 4) Signal2=TI3101, TI3102, TI3103, 6 1.644991e-03 0.9173634
 │  │  ├ 8) Signal1=TI3101, 2 4.314610e-07 0.8966558
 │  │  └ 9) Signal1=QI3105, TI3102 4 3.581534e-04 0.9277171
 │  └ 5) Signal2=PI3103, PI3104, PI3105, TI3104, TI3105 9 2.904812e-03 0.9808660
 │     ├ 10) Signal2=TI3104 3 469038e-04 0.9565722
 │     └ 11) Signal2=PI3103, PI3104, PI3105, TI3105 6 2.055057e-06 0.9930129
 └ 3) Signal1=FI3102, FI3103, FI3104, FI3105, i105, i106, PI3103,
    │             PI3104,PI3105,TI3103, TI3104 90 2.218348e-03 0.9921553
    ├ 6) Signal1=TI3103 2 6.312550e-04 0.97562671
    └ 7) Signal1=FI3102, FI3103, FI3104, FI3105, i105, i106, PI3103, PI3104, PI3105,
                   TI3104 88 1.003705e-03 0.9925391
```

5. Discussion

Step 1 of our method selected 15 beneficial signals of which the model score was 0.952, which is 0.017 higher than the base model score M_{B1} = 0.935, which evaluated all signals. Step 2 of our method selected 11 signals of which the model score was 0.988, which is 0.053 higher than the base model score M_{B1} = 0.935. Therefore, our method could improve the classification performance by 5.3%. A total of 22 signals were reduced to 11 signals. The reason why the signals of SG_1 in Figure 3, which had low factor loadings, are removed to use for the detection of abnormalities is because almost all of the displacements of the signals did not change between the normal and abnormal conditions. In contrast, the signals of SG_2 showed in Figure 4, which had high factor loadings, are beneficial for the detection of abnormalities because the displacements of the signals were separated for each normal and abnormal condition. We concluded that the factor analysis allows beneficial signals to be obtained easily and quickly.

Step 2 removed signals QI3105, TI3101, TI3102, and TI3103. The displacements of these signals are difficult for the detection of abnormalities because they might change very slowly from when the pipe blocking accident occurs Furthermore, the displacement of QI3105 might have returned to the initial level of displacement. This research revealed that the signals of which the displacement changes slowly or back to the source level of the displacement are not beneficial.

6. Conclusion

We proposed a method to select process signals to improve the classification performance of a diagnostic machine created from machine learning to detect abnormal operational conditions. Our method uses factor and decision-tree analyses to select signals and can be applied to various types of machine learning. In addition, our method has a mechanism to improve the performance considering combinations of signals. The evaluation results for our proposed method applied to a piping blockage accident in the flow process showed that the classification accuracy was improved by 5.3%. Our future work will evaluate whether or not our multi-agent diagnostic machine implemented with our optimised diagnostic machines is able to diagnose abnormalities individually and correctly.

Acknowledgements

This study includes some of the results of ‘Anomaly detection agents by advanced hybrid processing of Monju process data’ entrusted to Okayama University by the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT). The authors would like to thank Mr. Yuichiro Furuta who studied with me.

References

Ardi S., Minowa H. and Suzuki K., Detection of Runaway Reaction in a polyvinyl Chloride Batch Process Using Artificial Neural Networks, International Journal of Performability Engineering, Vol.5, No.4, pp.367-376, 2009.

Gofuku A., Takahashi M., Nagamatsu T., Mochizuki H, Furusawa H., Minowa H., Hybrid diagnostic agent system for the fast-breeder reactor “Monju”,International Journal of Nuclear Safety and Simulation(IJNS), Vol. 4, No.2, 2013.

Hiroyasu T., Nishioka M., Miki, M., Yokouchi H., 2009, Application of MOGA search strategy to SVM training data selection, 5th International Conference, EMO 2009, Nantes, 125-139.

Munesawa Y., Minowa H., Suzuki K., Development of fault diagnosis system using principal component analysis for intelligent operation support system, Chemical Engineering Transactions, Vol.31, pp.655-660, 2013, DOI: 10.3303/CET1331110.

Vapnik V., 1995, The Nature of Statistical Learning Theory. Springer-Verlag, New York, NY, United States.

Widodo A., Yang, B.-S., 2007, Support vector machine in machine condition monitoring and fault diagnosis, Mech. Syst. Signal Process. 21, 6, 2560-2574.

You C. H., Lee K. A., Li, H., 2009, An SVM kernel with GMM-supervector based on the Bhattacharyya distance for speaker recognition, IEEE Signal Process. Lett. 16, 1, 49-52.