Computers and Chemical Engineering 74 (2015) 59–65

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Input variable scaling for statistical modeling

Sanghong Kim[a,*], Manabu Kano[b], Hiroshi Nakagawa[c], Shinji Hasebe[a]

[a] Department of Chemical Engineering, Kyoto University, Kyoto 6158510, Japan
[b] Department of Systems Science, Kyoto University, Kyoto 6068501, Japan
[c] Formulation Technology Research Laboratories, Daiichi Sankyo Co., Ltd., Hiratsuka 2540014, Japan

ARTICLE INFO

Article history:
Received 8 July 2014
Received in revised form
22 December 2014
Accepted 29 December 2014
Available online 3 January 2015

Keywords:
Statistical model
Soft sensor
Input variable scaling
Pharmaceutical process
Distillation process

ABSTRACT

Input variable scaling is one of the most important steps in statistical modeling. However, it has not been actively investigated, and autoscaling is mostly used. This paper proposes two input variable scaling methods for improving the accuracy of soft sensors. One method statistically derives the input variable scaling factors; the other one uses spectroscopic data of a material whose content is estimated by the soft sensor. The proposed methods can determine the scales of the input variables based on their importance in output estimation. Thus, it can reduce the negative effects of input variables which are not related to an output variable. The effectiveness of the proposed methods was confirmed through a numerical example and industrial applications to a pharmaceutical and a distillation processes. In the industrial applications, the proposed methods improved the estimation accuracy by up to 63% compared to conventional methods such as autoscaling with input variable selection.

1. Introduction

In the process industry, one of the most important tasks is to ensure quality and to reduce operating cost. However, real-time measurement of product quality is not always available due to unacceptable measurement equipment cost and long measurement time. To solve this problem, research on soft sensors, which estimate product quality using real-time measurements, has been actively conducted (Kadlec et al., 2009; Kano and Fujiwara, 2013; Oh et al., 2013; Khatibisepehr et al., 2014). According to a questionnaire survey (Kano and Fujiwara, 2013), in 2009 soft sensors were working in over 400 distillation and chemical reaction processes at 15 companies in Japan. In addition, soft sensors have recently attracted much interest in the pharmaceutical industry to achieve a new quality assurance system composed of Quality by Design (QbD) and process analytical technology (PAT) (Roggo et al., 2007; Rajalahti and Kvalheim, 2011). Building a soft sensor requires many steps such as data acquisition, abnormal data detection, data preprocessing, input variable selection, model building, and model validation. Although input variable scaling, a data preprocessing method in which the values of each input variable are multiplied by the scaling factor of the input variable, can have significant effect on the estimation performance of soft sensors, research on input variable scaling has not been actively conducted. Hence, this paper focuses on input variable scaling, which is mathematically represented as

$$\tilde{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{\Lambda} \tag{1}$$

$$\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_M) \tag{2}$$

where $\boldsymbol{X} \in \Re^{N\times M}$ is the raw input variable matrix, in which the input variables are not scaled, $\tilde{\boldsymbol{X}} \in \Re^{N\times M}$ is the scaled input variable matrix, λ_m is a nonnegative input variable scaling factor for the m-th input variable, N is the number of samples, and M is the number of input variables. It is assumed that the mean of each input variable is zero without loss of generality. The input variable scaling affects important statistical properties of the data such as the distance between samples and the covariance of samples. It also affects the estimation result. For example, the m-th input variable x_m cannot have any influence on output estimation when λ_m is zero. Thus, $\boldsymbol{\Lambda} \in \Re^{M\times M}$ should be carefully selected to create accurate soft sensors.

In past research, autoscaling was commonly used (Engel et al., 2013; van den Berg et al., 2006; Todeschini et al., 1999). In addition, Pareto scaling, level scaling, Poisson scaling, range scaling, and

* Corresponding author. Tel.: +81 075 383 2677; fax: +81 075 383 2677.
E-mail address: kim@cheme.kyoto-u.ac.jp (S. Kim).

http://dx.doi.org/10.1016/j.compchemeng.2014.12.016
0098-1354/

VAST scaling (Keun et al., 2003) have been considered. The scaling factors in these methods are defined as

$$\frac{1}{\lambda_m} = \begin{cases} \sigma_m & \text{(autoscaling)} \\ \sqrt{\sigma_m} & \text{(pareto scaling)} \\ \bar{x}_m & \text{(level scaling)} \\ \sqrt{\bar{x}_m} & \text{(poisson scaling)} \\ x_{m,\max} - x_{m,\min} & \text{(range scaling)} \\ \dfrac{\sigma_m^2}{\bar{x}_m} & \text{(VAST scaling)} \end{cases} \tag{3}$$

where σ_m is the standard deviation of x_m, $\bar{x}_m$ is the mean value of x_m, $x_{m,\max}$ is the maximum value of x_m, and $x_{m,\min}$ is the minimum value of x_m. These methods define the input variable scaling factors based only on the information from the input variables such as their standard deviations and means. Hence, input variable scaling factors can be large for the input variables which are irrelevant to the output variable when these method are used, and the estimation performance of soft sensors may deteriorate. Some of the irrelevant input variables might be removed by using input variable selection methods such as the stepwise method (Hocking, 1976), variable influence on projection (VIP) (Wold et al., 2001) and least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996). It is, however, very difficult to remove all irrelevant input variables without removing any relevant input variables, and some irrelevant input variables generally remain after input variable selection. Thus, it is needed to determine the input variable scaling factors according to the importance of the input variables in output estimation. To take into account the importance of input variables in the output estimation, Kuzmanovski et al. (2009) used the genetic algorithm to optimize the input variable scaling factor. However, the computational burden of the genetic algorithm is considerable. Martens et al. (2003) proposed to use the magnitude of the undesired signals in measurements to determine the input variable scaling factors. But, this method is applicable only to spectroscopic data. To solve the above-mentioned problems, two input variable scaling methods are proposed. The proposed methods can determine the input variable scaling factors based on the importance of input variables in output estimation with short computational time. One of the proposed methods can be applied to any data.

2. Input variable scaling methods

Conventional input variable scaling methods such as autoscaling and range scaling do not determine the input variable scaling factors based on the importance of individual input variables in output estimation. These methods, therefore, can cause overfitting especially when the number of samples is small. One can reduce the effect of irrelevant input variables on output estimation by assigning small input variable scaling factors to those input variables. On the other hand, large input variable scaling factors should be assigned to input variables which have a large influence on an output variable.

We propose two methods to evaluate the influence of each input variable on an output variable and assign appropriate input variable scaling factors to input variables. The first one statistically derives the input variable scaling factors, while the second one uses spectroscopic data of a material whose content is estimated by a soft sensor.

2.1. Proposed method 1: data-based approach

Proposed method 1 statistically calculates the input variable scaling factor in an iterative manner. In this paper, the standardized regression coefficients of input variables in a partial least squares (PLS) model and the VIP scores are used as the input variable scaling factors, since they correlate to the importance of each input variable. The standardized regression coefficient is defined as the product of the regression coefficient β and the standard deviation σ of an input variable. The algorithm of proposed method 1 is as follows:

1. Prepare the raw input variable matrix $\boldsymbol{X}$ and an output variable vector $\boldsymbol{y} \in \Re^N$.
2. Set the iteration number i to 1 and the maximum iteration number to I.
3. Calculate the input variable scaling factor matrix $\boldsymbol{\Lambda}_0 = \mathrm{diag}(\lambda_{10}, \lambda_{20}, \ldots, \lambda_{M0})$ where λ_{m0} is $1/\sigma_{m0}$. Here, σ_{m0} is the standard deviation of the m-th input variable ($m = 1, 2, \ldots, M$) in the raw input variable matrix $\boldsymbol{X}$.
4. Let the scaled input matrix $\tilde{\boldsymbol{X}}_0 = \boldsymbol{X}\boldsymbol{\Lambda}_0$.
5. Calculate the new input variable scaling factor matrix

$$\boldsymbol{\Lambda}_i = \mathrm{diag}(\lambda_{1i}, \lambda_{2i}, \ldots, \lambda_{Mi}) \tag{4}$$

$$\lambda_{mi} = \begin{cases} |\beta_{mi}|\sigma_{mi} & \text{(standardized regression coefficient)} \\ \mathrm{VIP}_{mi} & \text{(VIP score)} \end{cases} \tag{5}$$

   for every m. Here, β_{mi}, σ_{mi} and VIP_{mi} denote the regression coefficient, the standard deviation and VIP score of the m-th input variable obtained using the scaled input matrix $\tilde{\boldsymbol{X}}_{i-1}$ and the output variable vector $\boldsymbol{y}$, respectively.
6. Calculate the new scaled input matrix $\tilde{\boldsymbol{X}}_i = \boldsymbol{X}\boldsymbol{\Lambda}_i$.
7. Finish the calculation if $i = I$. Otherwise set $i = i + 1$ and go to step 5.

Steps 3 and 4 in the above algorithm correspond to autoscaling. In step 5, the input variable scaling factors are updated, and the input variable matrix is updated in step 6. The convergence of this method is not guaranteed in all cases. However, the values of regression coefficients converged in most cases at least in the case studies conducted in this paper as shown in the next section.

The regression coefficient vector obtained by PLS is represented as

$$\boldsymbol{\beta}_{\mathrm{PLS}} = \boldsymbol{W}(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{W})^{-1}\boldsymbol{q} \tag{6}$$

$$\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_R] \tag{7}$$

$$\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_R] \tag{8}$$

$$\boldsymbol{q} = [q_1, q_2, \ldots, q_R]^{\mathrm{T}} \tag{9}$$

where $\boldsymbol{w}_r$, $\boldsymbol{p}_r$ and q_r are the weight vector, the loading vector of the input variable and the regression coefficient for the r-th latent variable (Kim et al., 2013).

The VIP score (Wold et al., 2001) of the m-th variable is defined as

$$\mathrm{VIP}_m = \sqrt{\frac{M\sum_{r=1}^{R}\left[(q_r^2\boldsymbol{t}_r^{\mathrm{T}}\boldsymbol{t}_r)\left(\frac{w_{mr}}{\|\boldsymbol{w}_r\|}\right)^2\right]}{\sum_{r=1}^{R}(q_r^2\boldsymbol{t}_r^{\mathrm{T}}\boldsymbol{t}_r)}} \tag{10}$$

where w_{mr} is the m-th component of the r-th weight vector $\boldsymbol{w}_r$. $\boldsymbol{t}_r$ is the r-th latent variable score.

2.2. Proposed method 2: knowledge-based approach

In the pharmaceutical and food industries, soft sensors are often used to estimate the content of an important material