Computers and Chemical Engineering 35 (2011) 2141–2151

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

Simultaneous data reconciliation and joint bias and leak estimation based on support vector regression

Yu Miao[a], Hongye Su[b,*], Wei Wang[a], Jian Chu[b]

[a] Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, Liaoning 116023, PR China
[b] National Key Laboratory of Industrial Control Technology, Institute of Cyber-Systems and Control, Department of Control Engineering, Yuquan Campus, Zhejiang University, Hangzhou 310027, PR China

ARTICLE INFO

Article history:
Received 12 March 2008
Received in revised form 20 May 2011
Accepted 3 June 2011
Available online 14 June 2011

Keywords:
Data reconciliation
Support vector regression
Parameter estimation
Gross error detection

ABSTRACT

Process data measurements are important for process monitoring, control, optimization, and management decision making. However, process data may be heavily deteriorated by measurement biases and process leaks. Therefore, it is significant to simultaneously estimate biases and leaks with data reconciliation. In this paper, a novel strategy based on support vector regression (SVR) is proposed to achieve simultaneous data reconciliation and joint bias and leak estimation in steady processes. Although the linear objective function of the SVR approach proposed is robust with little computational burden, it would not result in the maximum likelihood estimate. Therefore, to ensure accurate estimates, the maximum likelihood estimate is applied based on the result of the SVR approach. Simulation and comparison results of a linear recycle system and a nonlinear heat-exchange network demonstrate that the proposed strategy is effective to achieve data reconciliation and joint bias and leak estimation with superior performances.

1. Introduction

The measurement data of process variables usually contain random errors, which make the data not obey process constraints defined by the mass and energy balances. Therefore, data reconciliation should be applied to obtain accurate estimates of process variables based on the measurement data to support process control, optimization and monitoring well. However, data reconciliation results are probably corrupted heavily by the presence of gross errors like biases in measurements and leaks in process. Therefore, the effects of biases and leaks should be reduced or even eliminated if possible when applying data reconciliation. Many methods have been developed to address data reconciliation and gross error detection. The most widely used methods for data reconciliation and gross error detection are the global test (GT) (Reilly & Carpani, 1963), the measurement test (MT) (Mah & Tamhane, 1982), the nodal test (NT) (Reilly & Carpani, 1963), the generalized likelihood ratio (GLR) (Narasimhan & Mah, 1987) and the principal component test (PCT) (Tong & Crowe, 1995). Several strategies were developed to identify multiple biases, such as serial elimination (Rosenberg, Mah, & Iordache, 1987), serial compensation (Narasimhan & Mah, 1987), simultaneous or collective compensation (Kim, Kang, Park, & Edgar, 1997). Some new methods have been applied to real industrial processes, which range from statistical test methods to robust statistics methods (Arora & Biegler, 2001; Wang & Romagnoli, 2003; Wongrat, Srinophakun, & Srinophakun, 2005), from sequential or combinatorial methods to simultaneous data reconciliation and gross error detection methods. However, few methods could achieve simultaneous data reconciliation and joint bias and leak estimation, such as GLR and the simultaneous estimation of gross errors (SEGE) method (Sanchez, Romagnoli, Jiang, & Bagajewicz, 1999). Although the latter one usually has a better performance, it will be too much computational overhead in a large scale system as it will test too many possible combinations of suspected variables, and the test procedure of which is only suitable for linear systems.

Simultaneous data reconciliation and gross error estimation can be addressed as a model identification and parameter estimation problem, and the Akaike information criterion (AIC) (Akaike, 1974) has been applied, which is supported by an earlier work (Yamamura, Nakajima, & Matsuyama, 1988), where AIC is applied to identify biased measurements in a least squares framework for gross error detection. Due to the combinatorial nature of the problem attempted, a branch and bound method is suggested to solve the problem. There also exists some related work. A mixed integer linear program (MILP) approach has been presented (Soderstrom, Edgar, Russo, & Young, 2000), which is similar to AIC in the form. However, this approach is computationally expensive as it requires

* Corresponding author. Tel.: +86 0571 8795 1200; fax: +86 0571 8666 7616.
E-mail addresses: hysu@iipc.zju.edu.cn, suhy@supcon.com (H. Su).

0098-1354/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compchemeng.2011.06.002

a discrete decision with two binary variables for each measurement. Arora and Biegler (2001) argued that the mixed integer non-linear program (MINLP) approach is a direct minimizer of AIC. However, the fixed penalty coefficient in AIC makes it inappropriate to penalize detecting too many gross errors especially in a small example usually encountered in data reconciliation problem. Furthermore, these approaches have not yet addressed process leaks. Similar to AIC, there are some other criterions such as Bayesian information criterion (BIC), whose penalty coefficient changes with the sample size. Although the penalty for additional parameter is stronger than that of AIC, BIC still cannot solve the problems encountered by AIC well for data reconciliation, such as computational burden and process leaks. Therefore, a novel approach is needed, which should be more effective to address joint bias and leak estimation with robustness and little computational load.

The support vector algorithm is a nonlinear generalization of the Generalized Portrait algorithm developed in Russia in the sixties. As such, it is firmly grounded in the framework of statistical learning theory, or Vapnik–Chervonenkis theory (known as VC theory), which has been developed over the last three decades by Vapnik, Chervonenkis and others. According to statistical learning theory, minimizing empirical risk, which will lead to overfitting and thus bad generalization properties, is replaced by minimizing regularized risk with adding a capacity control term to objective function. Recently, the support vector regression (SVR) approach has been introduced to address simultaneous data reconciliation and measurement biases detection problem (Miao, Su, & Chu, 2009; Miao, Su, Xu, & Chu, 2009) benefiting from its robust and excellent nature of classification to efficiently distinguish biased measurements from normal ones. In this paper, with considering process model with leaks and its complexity, SVR approach is extended to deal with simultaneous data reconciliation and joint bias and leak estimation. In order to reduce the computational load led by MINLP and to increase robustness to bias and leak, a simplified linear objective function is used in the proposed SVR method according to a basic SVR algorithm named ε-SVM (Scholkopf, Smola, Williamson, & Bartlett, 2000). Although the linear objective function is robust to biases and leaks with less computational burden, it will not result in the maximum likelihood estimate. Therefore, the maximum likelihood estimate is applied based on the result of the SVR approach to guarantee accurate estimates. Meanwhile, it can be easily revealed that the combinatorial search and test of measurement biases and process leaks in the SEGE approach can be included in the framework of our approach. Therefore, our approach, like the SEGE, could give results that are consistent with the gross error equivalency theory (Bagajewicz & Jiang, 1998). Meanwhile, the MILP approach could be included in the proposed SVR approach.

In this paper, a linear recycle system and a nonlinear heat-exchange network are used for case study. In the linear recycle system case study, the effect of penalty on the power of detecting biases and leaks for data reconciliation are studied on AIC, BIC and the SVR approach, and a suggestion of penalty tuning procedure for the SVR approach is made. Then the performances of the SVR approach are demonstrated and compared with those of MILP approach and the extended MINLP approach based on AIC on the linear recycle system. In the second case study, we try to solve a nonlinear data reconciliation problem with process leaks and measurement biases on a heat-exchange network, and compare the performances of the SVR approach with a robust data reconciliation approach based on generalized T (GT) distribution to demonstrate the efficiency of our approach on nonlinear system. The simulation and comparison results in both case studies show that the proposed method based on SVR approach is robust, effective and accurate for data reconciliation with joint measurement bias and process leak estimation.

2. Mixed integer program approaches with AIC for data reconciliation

The general form of simultaneous data reconciliation and gross error detection can be described as following,

$$\min_{x,u,p} F(x^{\mathrm{M}}, x) \quad \text{s.t.} \quad \begin{aligned} & h(x,u,p)=0 \\ & x_i^{\mathrm{L}} \le x_i \le x_i^{\mathrm{U}} \\ & p_i^{\mathrm{L}} \le p_i \le p_i^{\mathrm{U}} \\ & u_i^{\mathrm{L}} \le u_i \le u_i^{\mathrm{U}} \end{aligned} \tag{1}$$

where $F(\cdot)$ is the objective function, x^{M} is the set of measurement data of the corresponding variable set x, p is the set of parameters, u is the set of unmeasured variables, $h(\cdot)$ is the set of equality model constraints, the subscript i denotes the ith element in the corresponding set, the superscripts L and U denote the lower and upper bounds of the corresponding variables, respectively.

In the formula given above, data reconciliation and bias detection can be addressed as a model identification and parameter estimation problem, where estimated parameters p could be seen as measurement biases and process leaks. Meanwhile, it is usually a small example problem as shown in (1) because there is usually only one set of measurements obtained for data reconciliation. If more than one model could be fitted to the data set, it is necessary to identify the best model and its parameters by suitable model evaluation criteria. AIC has been used for this purpose, which takes the form of a penalized likelihood. It is given by,

$$\mathrm{AIC} = -2\log(L(\hat{\theta})) + 2k \tag{2}$$

where $L(\hat{\theta})$ is the maximized likelihood function, and k is the number of free parameters in the model. The model with minimum AIC value is chosen as the best model to be used.

Based on the assumption that the random errors possess a normal distribution after removing the gross errors, Yamamura et al. (1988) first introduced AIC into data reconciliation and gross error detection problem for a linear system. Through dividing the set of measurement values into sets with gross errors and without gross error, a branch-and-bound strategy was proposed to solve the problem. The procedure of the branch-and-bound strategy for linear system was translated by Arora and Biegler (2001) into a MINLP with binary variables identifying the variables with gross errors, the formula of which is as the following,

$$\min_{x_i,\mu_i,r_i,z_i} \sum_{i=1}^{n} \left[\frac{(x_i^{\mathrm{M}} - x_i)}{\sigma_i} - \frac{\mu_i}{\sigma_i} \right]^2 + 2\sum_{i=1}^{n} r_i$$
$$\text{s.t.} \quad \begin{aligned} & \boldsymbol{Ax} = 0 \\ & \mu_i \le U_i r_i \\ & -\mu_i \le U_i r_i \\ & \mu_i - z_i U_i - z_i L_i + L_i r_i \le 0 \\ & -\mu_i + z_i U_i + z_i L_i + L_i r_i \le L_i + U_i \\ & z_i \le r_i \\ & 0 \le x_i \le X_i \\ & r_i, z_i \in \{0, 1\} \end{aligned} \tag{3}$$

where n is the number of measured variables, x_i is the ith variable in the measured variable vector $\boldsymbol{x}$, x_i^{M} is the measurement of the ith variable, σ_i is the standard deviation of the ith measurement, μ_i is the magnitude of bias in the ith measurement, r_i is a binary variable denoting existence of bias in the ith measurement, z_i is a binary variable for the sign of the ith bias, $\boldsymbol{A}$ is the matrix for linear