Article

Missing Data Probability Estimation-Based Bayesian Outlier Detection for Plant-Wide Processes with Multisampling Rates

Ying Tian ^{1,*} , Zhong Yin ¹ and Miao Huang ²¹ School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China; yinzhong@usst.edu.cn² Ningbo Institute of Technology, Zhejiang University, Ningbo 315100, China; huangmiao_2010@126.com

* Correspondence: tianying@usst.edu.cn; Tel.: +86-131-2097-5286

Received: 16 September 2018; Accepted: 1 October 2018; Published: 10 October 2018

Abstract: Traditional outlier detection methods assume that the sampling time and interval are the same. However, for plant-wide processes, since the signal change rate of different devices may vary by several orders of magnitude, the measured data in real-world systems usually have different sampling rates, resulting in missing data. To achieve reliable outlier detection, a missing data probability estimation-based Bayesian outlier detection method is adopted. In this strategy, the expectation–maximization (EM) algorithm is first used to estimate the likelihood probability of different evidence under different process statuses by using the history dataset which contains complete and incomplete samplings. Secondly, the realization of unavailable parts in the monitoring point is estimated as a probability through historical data and online moving horizon data. Bayesian theory and likelihood probability are then used to calculate the outlier posterior probability of different realization. Finally, the outlier probability of the monitoring sampling is calculated by the probability of different realizations and the corresponding outlier probability. Using the Tennessee Eastman (TE) dataset, a simulation indicates that the proposed method exhibits a significant improvement over the complete data method.

Keywords: outlier detection; missing data estimation; Bayesian; expectation–maximization (EM); multisampling rates; plant-wide process

1. Introduction

Outlier detection, an important research topic in data mining, has attracted wide attention in academic and applied fields. Outliers appear in the data, usually as a result of process disturbance or instrument drift, which may dramatically undermine subsequent analysis based on the data. Therefore, with the development of science and technology, many outlier detection methods have been proposed; examples include a distance-based outlier detection method for multidimensional datasets [1]; a fuzzy rough semi-supervised outlier detection approach with the help of labeled samples [2]; neural networks-based abnormal detection [3,4]; fuzzy discrimination extended and applied to outlier detection [5,6]; rough set-based attribute selection to handle outlier detection for high dimension data [7]; rough entropy used to calculate the degree of outliers [8]; and, ellipsoidal support vector machine-based outlier detection [9].

Traditional outlier detection methods usually assume that the sampling time and interval are the same. However, for complex multivariable control systems, the signal change rate of different devices is different. Under this situation, if a short sampling period is adopted, the cost of the control system is improved, while if a long period is used, then important information contained in the fast-changing

variables will be ignored. In other words, it is impossible to use a single sampling interval in a complex system. In fact, most real-world systems feature multisampling rates. Since a subset of the variables is sampled at a higher rate and another subset of the variables is sampled at a lower rate, the dataset collected from the system will be affected by a problem of incomplete data. For the incomplete dataset, the deleting method is first proposed, in which the records with missing data are deleted directly to obtain a complete dataset. Although this method is easy to implement, a significant amount of effective information will be deleted if the amount of missing data is large, and the real-time accuracy of outlier detection will be affected if only the complete sampling is used for modeling and detection. Therefore, imputation methods are used to handle incomplete data, such as mean completer, combinatorial completer, regression, hot deck, and multiple imputation. These methods achieve dataset integrity by estimating the value of missing data [10–13]. In addition, to make the best use of existing data without adding or deleting data, methods of data mining based on the missing dataset have also been proposed, for example, the artificial neural network method [14,15], or the rough set reasoning method [16,17].

Despite great achievements having been made for handling missing data and detection of outliers, there are still some shortcomings. For example, because all of these methods assign a definite value to missing data and classify a sample as being normal or an outlier directly, a wrong imputation value or wrong classification will significantly affect subsequent analysis and processing. To avoid such errors, probabilistic estimation is a good alternative. As an efficient tool in probabilistic inference, the Bayesian method has been developed, for example, a Bayesian probabilistic approach has been proposed to find the internal relations and to identify the faults possibly present in the system given the current observations [18], and a novel Bayesian probabilistic diagnostic framework for control loop monitoring has been established and demonstrated [19]. However, the traditional Bayesian method still requires that all monitor readings be available at the same time. To handle the missing data problem, a method based on marginalization over an underlying complete evidence matrix has been proposed to circumvent missing data problems and to realize probability estimation for unmeasured data [20]. While this method focuses only on the case where the missing data follow one specific pattern, in order to deal with the problem of multiple missing data patterns, [21] proposed an expectation–maximization (EM) approach to estimate the likelihood probability of different realizations. Based on the EM approach, considering the plant-wide process problem as well as the asynchronous measurements question, [22] proposed a Bayesian marginalization method within a moving horizon for online incomplete measurements and established a Bayesian diagnosis system revealing both the underlying fault status of the whole plant and the unavailable statuses of the corresponding local units.

Considering the missing data problem of outlier detection, based on the Bayesian probability framework and the marginalization method, as well as the EM algorithm, a missing data estimation and outlier detection strategy is proposed. The contribution contains four aspects:

- (1) Probability estimation for the realization of unavailable parts in the monitoring sampling is executed through a marginalization method over all historical data and online moving horizon data, including complete data and incomplete data.

- (2) The EM algorithm for missing data estimation is used to calculate the likelihood probability of different realizations under different process statuses.

- (3) The outlier probability of different realizations is obtained through Bayesian theory.

- (4) The total outlier probability of the current incomplete sampling is calculated using the full probability theory with the probability of different realizations and its corresponding outlier probability.

The innovation is that the method proposed in this study is more applicable to real-world processes. It is well known that most real-world complex multivariable processes feature multisampling rates, resulting in a monitoring dataset with missing data. The question of how this missing data is handled is important. The proposed strategy can make full use of the information contained in historical incomplete data and online incomplete data, to determine whether current data, especially current

incomplete data, are outliers or not. Incomplete data used in the modeling stage makes detection more accurate, and the use of online incomplete data enables more timely detection.

The remainder of this paper is structured as follows: first, the problem statement and motivation analysis are briefly reviewed. Then, the Bayesian outlier detection method for processes with multisampling rates is described in detail. The efficiency of the proposed approach is illustrated by the TE process. Finally, conclusions and perspectives are provided.

2. Problem Statement and Motivation Analysis

A data acquisition system (DAS) is a complex system composed of computer control technology, a computer network, and intelligent instruments. The widely used DAS in a plant-wide process is a distributed structure with an upper system, several data collection sites, and communication lines (shown in Figure 1). Usually, in this system, the sampling time and interval are assumed to be the same. In a complex multivariable plant-wide process, however, given the complexity of the controlled and monitored object, the signal change rate of different devices differs, so the sampling period of the related detection device is different, for example, the change rate of a temperature signal and an electrical signal may vary by several orders of magnitude. Although a short sampling period can be used to achieve better outlier detection, this will require an increased cost of the computer control system. Thus, it is impossible to use a single sampling period in all parts of the system. To this end, the best method is to adopt different sampling periods for different change rate signals.

Figure 1. Data acquisition system.

When multisampling rates are adopted, the dataset collected from the system will be affected by the problem of missing data. For instance, for a plant-wide monitoring process, suppose that the input is $\begin{bmatrix} \pi_1 & \pi_2 & \pi_3 \end{bmatrix}$, $\pi_i \in \{0, 1\}$, and the sampling interval for π_1 is T , for π_2 is $2T$, and for π_3 is $3T$. If the symbol “*” is used to represent unavailable values, then the monitoring dataset is illustrated in Table 1. This shows that there is only one complete data point in every six samplings. Below we discuss how to realize outlier detection for this kind of dataset.

Table 1. Monitoring dataset with missing value.

Sample	$\pi_1(T)$	$\pi_2(2T)$	$\pi_3(3T)$	Possible Realization
d^1	0	*	*	$[0 \ 0 \ 0]$, $[0 \ 0 \ 1]$, $[0 \ 1 \ 0]$, $[0 \ 1 \ 1]$
d^2	0	1	*	$[0 \ 1 \ 0]$, $[0 \ 1 \ 1]$
d^3	0	*	0	$[0 \ 1 \ 0]$, $[0 \ 0 \ 0]$
d^4	0	1	*	$[0 \ 1 \ 0]$, $[0 \ 1 \ 1]$
d^5	0	*	*	$[0 \ 0 \ 0]$, $[0 \ 0 \ 1]$, $[0 \ 1 \ 0]$, $[0 \ 1 \ 1]$
d^6	0	1	0	$[0 \ 1 \ 0]$

The conventional outlier detection method requires the sampling interval and time to be consistent. Therefore, to realize monitoring for a multisampling rate system, the missing data should be handled first. The traditional methods mainly include deletion and imputation. The deleting method directly removes incomplete data; for the above example, only one complete sampling data point can be used in every six monitoring samples, which leads to the loss of a significant amount of effective information and to monitoring delay. To use as much information as possible, the imputation method, which gives a reasonable substitute value for missing data, is much more suitable. Through imputation, a complete dataset can be constructed. For the above example, the possible realizations for each incomplete data point are shown in Table 1; however, we need to determine which realization is most likely. In other words, the key question is how to derive imputation values that are as close to the missing original data as possible to reduce the estimated error. That is, the objective of our work is to provide an estimation method for missing data and an outlier detection method for multisampling rates in a plant-wide process. There are several problems to be considered:

(1) Since most methods provide definite estimated values for missing data, wrong imputation will significantly affect subsequent analysis and processing. Can probabilistic estimation be considered for missing data realization estimation in order to avoid such errors?

(2) If probabilistic estimation is adopted for the realization of current incomplete monitoring samples, then a follow-up issue is how to calculate outlier probability for each realization and how to classify an incomplete sample through the realization probability and the outlier probability of each realization.

(3) Given the complexities of probabilistic estimation, and the large number of variables in a plant-wide process system, if all variables with continuous values are used in the monitoring index, then the computing complexity will exceed the ability of a general computer. Thus, determining how to form an effective monitoring index is also an urgent problem.

All of these problems will be effectively solved in the next section.

3. Bayesian Outlier Detection for Plant-Wide Processes with Multisampling Rates

Considering the shortcomings of traditional methods, based on the multisampling rates of plant-wide processes, missing data probability estimation-based Bayesian outlier detection is adopted here. In this strategy, considering the computing complexity of plant-wide processes, given that both the historical data and online horizon data are multisampling rates with incomplete data, the research includes four aspects: (1) to reduce complexity, variables with the same sampling period are placed in a sub-block, and PCA is performed for each sub-block to form monitoring evidence; (2) marginalization-based probability estimation for realization of current incomplete evidence is executed through historical multisampling rate samples and online moving horizon data; (3) the EM algorithm is used to estimate the likelihood of different evidence using multisampling rate historical data; and (4) the posterior probability of different process statuses for current incomplete data is calculated according to Bayesian theory and full probability theory.

3.1. Marginalization-Based Realization Estimation

For the multisampling rates system, in order to ensure the total amount of modeling and monitoring data, this study intends to probabilistically estimate the realization of missing data by using complete and incomplete historical data, as well as online moving windows. Several variables need to be defined before the estimation is made:

(1) Evidence, E , which is usually the monitoring variable of a process. For a system with B monitors, its evidence can be expressed as $E = \{\pi_1, \pi_2, \dots, \pi_B\}$, where π_i is the i th source with q_i discrete values. Therefore, the collection of all possible evidence is $\varepsilon = \{e_1, e_2, \dots, e_K\}$, where $K = \prod_{i=1}^B q_i$. However, for a plant-wide system, the variable's measured values are continuous, and the number of process variables is large; thus, taking each process variable as a source of evidence may be beyond the capability of a normal computer. Therefore, the suitable evidence should be designed first in this study. To reduce its size, the multi-block method is adopted to handle data and obtain suitable evidence. First, variables with the same sampling rate are placed into a sub-block, a PCA model is established for each sub-block, and the T^2 and SPE statistics and control limits of each sub-block can be obtained. For PCA details, please refer to [23].

According to the statistic and control limit, the evidence can be generated as

$$\begin{cases} \pi_i = 0 & \text{if } T_i^2 \leq T_{i,\text{lim}}^2 \text{ and } SPE_i \leq SPE_{i,\text{lim}} \\ \pi_i = 1 & \text{if } T_i^2 > T_{i,\text{lim}}^2 \text{ or } SPE_i > SPE_{i,\text{lim}} \end{cases} \quad (1)$$

where $i = 1, 2, \dots, m$ denotes the block number; π_i is the i th source in evidence; T_i^2 and $T_{i,\text{lim}}^2$ are the T^2 statistic and T^2 control limit for the i th sub-block; SPE_i and $SPE_{i,\text{lim}}$ are the SPE statistic and SPE control limit for the i th sub-block; and $\pi_i = 1$ indicates that the data of the i th sub-block is an outlier.

Note that the number of principal components is selected by the cumulative percent variance (CPV). For sub-block data, the covariance matrix of the data is calculated and related eigenvalues are sorted. The ratio between the first k eigenvalues and the sum of all eigenvalues is defined as the CPV, which is used to represent the proportion of data explained by the first k principal components in all data. A reasonable k is very important for PCA modeling. In this research, we choose the first k principal components whose CPV is greater than 85% as the modeling principal components, as shown in the equation $\sum_{i=1}^k \lambda_i / \sum_{i=1}^m \lambda_i > 85\%$.

(2) Process status, S , which is the internal state of the system. A system with G possible internal states is denoted $S = \{S_1, S_2, \dots, S_G\}$. For the outlier detection problem, $G = 2$.

(3) History dataset, D , which is labeled data. A historical dataset with N historical training samples can be expressed as $D = \{d^1, d^2, \dots, d^N\}$, where d^t contains evidence e^t and process internal status S^t at time t : $d^t = \{e^t, S^t\}$.

(4) Online horizon, H , which is defined as $H = \{h^0, h^1, h^2, \dots, h^{r-1}\}$, where h^l refers to the l th forward sample from the current sample, and r is the length of the horizon. Data h^l consists of an evidence vector e^l and posteriori probability of F_j under evidence e^l .

Next, for a current evidence sample, h , with missing data, a marginalization-based solution is used for its realization probability estimation through its observed part, y_h , the history training data, and online moving horizon data, which can be expressed as

$$p(e_i|y_h, H, D) = \int_{\Omega} p(e_i|y_h, \Psi, H, D) f(\Psi|y_h, H, D) d\Psi \quad (2)$$

where $\Psi = \{\psi_{1|y_h}, \psi_{2|y_h}, \dots, \psi_{K_h|y_h}\}$, $\psi_{i|y_h} = p(e_i|y_h, H, D)$, K_h is the number of the possible realizations of y_h and Ω is the space of all possible parameters in Ψ . For instance, for a three-dimensional system with two possible discrete values, assuming the current incomplete evidence

is $\begin{bmatrix} 1 & * & * \end{bmatrix}$, then its possible evidence set $R_h = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$, and the number of the possible realizations, K_h , is 4.

In Equation (2), $f(\Psi|y_h, H, D)$ can be calculated by the Bayes' rule as

$$f(\Psi|y_h, H, D) = \frac{p(H|\Psi, y_h, D)f(\Psi|y_h, D)}{p(H|y_h, D)} \quad (3)$$

where the numerator can be calculated by integration over the likelihood space Ω .

$$p(H|y_h, D) = \int_{\Omega} p(H|\Psi, y_h, D)f(\Psi|y_h, D) \quad (4)$$

The Dirichlet distribution with Dirichlet parameters is commonly used to estimate the probability of Ψ

$$f(\Psi|y_h, D) = \frac{\Gamma\left(\sum_{i=1}^{K_h} \alpha(e_i|y_h, D)\right)}{\prod_{i=1}^{K_h} \alpha(e_i|y_h, D)} \prod_{i=1}^{K_h} (\psi_{i|y_h})^{\alpha(e_i|y_h, D)-1} \quad (5)$$

where $\Gamma(\cdot)$ is the Gamma function, $\alpha(e_1|y_h, D), \alpha(e_2|y_h, D), \dots, \alpha(e_{K_h}|y_h, D)$ and $\alpha(e_i|y_h, D)$ are the number of prior samples for the possible realization, e_i , given the observed part, y_h , which is calculated through the historical training data, D .

The likelihood of the samples that are possible realizations of h in the horizon can be written as

$$p(H|\Psi, y_h, D) = \prod_{k=1}^{N_{Hy_h}} p(h_{y_h}^k | \Psi, y_h, D) = \prod_{i=1}^{K_h} (\psi_{i|y_h})^{n(e_i|y_h, H)} \quad (6)$$

where N_{Hy_h} is the expected number of samples that are possible realizations of h in the horizon, and $n(e_i|y_h, H)$ is the number of e_i in the moving horizon.

Taking Equations (3)–(6) into Equation (2)

$$p(e_i|y_h, H, D) = \int_{\Omega} \psi_{i/y_h} \frac{\prod_{i=1}^{K_h} (\psi_{i|y_h})^{n(e_i|y_h, H)} \frac{\Gamma\left(\sum_{i=1}^{K_h} \alpha(e_i|y_h, D)\right)}{\prod_{i=1}^{K_h} \alpha(e_i|y_h, D)} \prod_{i=1}^{K_h} (\psi_{i|y_h})^{\alpha(e_i|y_h, D)-1}}{\int_{\Omega} \prod_{i=1}^{K_h} (\psi_{i|y_h})^{n(e_i|y_h, H)} \frac{\Gamma\left(\sum_{i=1}^{K_h} \alpha(e_i|y_h, D)\right)}{\prod_{i=1}^{K_h} \alpha(e_i|y_h, D)} \prod_{i=1}^{K_h} (\psi_{i|y_h})^{\alpha(e_i|y_h, D)-1}} d\Psi \quad (7)$$

Using the same derivation procedures as in [20], the realization probabilities can be achieved as [22]

$$p(e_i|y_h, H, D) = \frac{n(e_i|y_h, H) + \alpha(e_i|y_h, D)}{N_{Hy_h} + N_{Dy_h}} \quad (8)$$

where N_{Dy_h} is the total number of samples that are possible realizations of h in historical data.

Generally, to reflect the prior knowledge, a number of prior samples should be added to the history data; that is $\alpha(e_i|y_h, D) = n(e_i|y_h, D) + \alpha(e_i|y_h)$, where $\alpha(e_i|y_h)$ is the number of prior hypothetical samples for the e_i that are possible realizations of h , and $n(e_i|y_h, D)$ is the number of e_i that are possible realizations of h in historical data.

Then, the realization probability is re-expressed as

$$p(e_i|y_h, H, D) = \frac{n(e_i|y_h, H) + n(e_i|y_h, D) + \alpha(e_i|y_h)}{N_{Hy_h} + N_{Dy_h} + A_{y_h}} \quad (9)$$

where A_{y_h} are the total prior samples that are possible realizations of h .

To calculate the realization probability in Equation (9), the below should be performed

$$n(e_i|y_h, D) = \sum_{j=1}^G N_{S_j} \cdot p(e_i|S_j, D) \quad (10)$$

$$N_{D_{y_h}} = \sum_{i=1}^{K_h} n(e_i|y_h, D) \quad (11)$$

where N_{S_j} is the number of samples with process status S_j in historical dataset D , $p(e_i|S_j, D)$ is the likelihood which will be introduced in the next section, and $e_i \in R_h$.

For the priori information, the uniform prior is employed, therefore

$$\alpha(e_i|y_h) = 1 \quad (12)$$

$$A_{y_h} = K_h \quad (13)$$

where $e_i \in R_h$.

Then

$$n(e_i|y_h, H) = \sum_{k=1}^r n(e_i|h^k, y_h, H) \quad (14)$$

$$n(e_i|h^k, y_h, H) = \begin{cases} p^*(e_i|h^k, y_h, H), & \text{if } e_i \in R_{h^k} \\ 0, & \text{if } e_i \notin R_{h^k} \end{cases} \quad (15)$$

$$N_{H_{y_h}} = \sum_{i=1}^{K_h} n(e_i|y_h, H) \quad (16)$$

where h^k is the sample in the online moving horizon, r is the length of the horizon, $p^*(e_i|h^k, y_h, H)$ is the recorded realization probability of sample h^k , and $e_i \in R_h$.

In summary, the realization probability is calculated through historical data, online horizon data, and prior knowledge.

3.2. Expectation–Maximization-Based Likelihood Probability Estimation

Here, we introduce the estimation method for the likelihood probability of different evidence under different process statuses. Considering that the historical data is multisampling rates with missing data, the expectation–maximization (EM) method for multiple missing data patterns proposed in [21] is adopted for likelihood estimation here.

First of all, the EM algorithm with missing data is introduced, which iteratively switches between the expectation step (E-step) and maximization step (M-step) to find the maximum likelihood estimate of parameters of interest.

In the E-step, the expected value of the log-likelihood function (Q-function) is built by using the previously estimated parameter, Θ^{old}

$$Q(\Theta|\Theta^{old}) = E_{C_{miss}|C_{obs}, \Theta^{old}} [\log(p(C_{miss}|C_{obs}, \Theta))] \quad (17)$$

where Θ is the parameter set to be estimated; Θ^{old} is the estimation result in the previous step; C_{miss} is the unobserved data; and C_{obs} is the observed dataset.

In the M-step, the new estimation of the parameter set is obtained by maximizing the Q-function obtained from the E-step

$$\Theta^{new} = \underset{\Theta}{\operatorname{argmax}} Q(\Theta|\Theta^{old}) \quad (18)$$

This iteration continues until some stop criterion is satisfied.

Next, we describe how to use the EM algorithm to solve the outlier identification problem. The likelihood probability is denoted $\theta_k = p(e_k|S_j, D)$, which is interpreted as the probability of e_k

under process status S_j , and $\Theta = [\theta_1, \theta_2, \dots, \theta_K]$ is the likelihood probability set for process status S_j . As for the outlier detection problem, there are two process statuses: normal and outlier. The optimized parameter set of all process statuses is $\Xi = [\Theta_1^{opt}, \Theta_2^{opt}]$.

The likelihood probability set $\Theta = [\theta_1, \theta_2, \dots, \theta_K]$ for S_j is estimated first. Since the process involves multisampling rates, the monitoring data subset D_{S_j} of S_j contains the complete part D_c and the incomplete part D_{ic} ; i.e., $D_{S_j} = \{D_c, D_{ic}\}$, and

$$D_c = \{d_c^1, d_c^2, \dots, d_c^{N_c}\} \quad (19)$$

$$D_{ic} = \{d_{ic}^1, d_{ic}^2, \dots, d_{ic}^{N_{ic}}\} \quad (20)$$

where d_c^i is the complete evidence, N_c is the total number of complete evidence, d_{ic}^i is the incomplete evidence, and N_{ic} is the number of incomplete evidence in D_{S_j} .

Given that the data are independent, the probabilities of the complete part and the incomplete part under current likelihood probability parameter set Θ^{old} can be expressed, respectively

$$p(D_{ic}|\Theta^{old}) = \prod_{t=1}^{N_{ic}} p(d_{ic}^t|\Theta^{old}) \quad (21)$$

$$p(D_c|\Theta^{old}) = \prod_{t=1}^{N_c} p(d_c^t|\Theta^{old}) = \prod_{k=1}^K \theta_k^{n(e_k|D_c)} \quad (22)$$

The total likelihood function for D_{S_j} can be calculated as

$$\begin{aligned} L(D_{S_j}|\Theta^{old}) &= \log p(D_c, D_{ic}|\Theta^{old}) \\ &= \log [p(D_c|\Theta^{old}) p(D_{ic}|\Theta^{old})] \\ &= \log \left[\prod_{k=1}^K \theta_k^{n(e_k|D_c)} \prod_{t=1}^{N_{ic}} p(d_{ic}^t|\Theta^{old}) \right] \end{aligned} \quad (23)$$

Moreover, since the incomplete data entries of D_{ic} can be further partitioned into the monitoring part, $y = \{y_1, y_2, \dots, y_{N_{ic}}\}$, and the missing part, $z = \{z_1, z_2, \dots, z_{N_{ic}}\}$, we have

$$L(D|\Theta^{old}) = \sum_{k=1}^K n(e_k|D_c) \log \theta_k + \sum_{t=1}^{N_{ic}} \log p(z_t, y_t|\Theta^{old}) \quad (24)$$

Taking Equation (24) into Equation (17), the Q-function is denoted as

$$\begin{aligned} Q(\Theta|\Theta^{old}) &= \sum_{\mathbf{Z}} p(\mathbf{z}|\mathbf{y}, D_c, \Theta^{old}) \log p(D_{ic}, D_c|\Theta) \\ &= \sum_{\mathbf{Z}} p(\mathbf{z}|\mathbf{y}, D_c, \Theta^{old}) \left(\sum_{k=1}^K n(e_k|D_c) \log \theta_k + \sum_{t=1}^{N_{ic}} \log p(z_t, y_t|\Theta) \right) \end{aligned} \quad (25)$$

where $\mathbf{Z}$ is the space for all possible values of the realization $\mathbf{z}$.

Following the derivations of [21], the Q-function can be expressed as

$$Q(\Theta|\Theta^{old}) = [n(\varepsilon|D_c) + n(\varepsilon|D_{ic}, \Theta^{old})]^T \log \Theta \quad (26)$$

where $\varepsilon = [e_1, e_2, \dots, e_K]^T$, and $n(e_i|D_c)$ is the number of evidence $e = e_i$ in the complete dataset, and $n(e_i|D_{ic}, \Theta^{old})$ is the estimated amount of evidence $e = e_i$ in the incomplete dataset under the likelihood parameter set Θ^{old} .

Considering that $\theta_k = 1 - \bar{\theta}_k = 1 - \sum_{j \neq k} \theta_j$, Equation (26) can be expanded as

$$Q(\theta_k | \Theta^{old}) = \left[n(e_k | D_c) + n(e_k | D_{ic}, \Theta^{old}) \right] \cdot \log \theta_k + \left[n(\bar{e}_k | D_c) + n(\bar{e}_k | D_{ic}, \Theta^{old}) \right] \cdot \log(1 - \theta_k) \quad (27)$$

where

$$\begin{aligned} n(e_k | D_{ic}, \Theta^{old}) &= \sum_{t=1}^{N_{ic}} p(e_k | d_{ic}^t, \Theta^{old}) \\ n(\bar{e}_k | D_{ic}, \Theta^{old}) &= \sum_{j \neq k} n(e_j | D_{ic}, \Theta^{old}) \\ p(e_k | d_{ic}^t, \Theta^{old}) &= \begin{cases} \frac{p(e_k | \Theta^{old})}{\sum_{e_j \in d_{ic}^t} p(e_j | \Theta^{old})}, & e_k \in d_{ic}^t \\ 0, & e_k \notin d_{ic}^t \end{cases} \end{aligned} \quad (28)$$

By taking the first derivative with respect to θ_k and setting it to zero to obtain the maximum value of the Q-function, the estimation of θ_k is achieved as

$$\theta_k^{new} = \frac{n(e_k | D_c) + n(e_k | D_{ic}, \Theta^{old})}{N_c + N_{ic}} \quad (29)$$

Moreover, the initial conditions are set through the available complete evidence, $\theta_k^0 = \frac{n(e_k | D_c)}{N_c}$. Finally, the process is repeated until the parameters converge.

3.3. Bayesian and Full Probability-Based Outlier Detection

Based on the likelihood probability of each evidence under different process statuses, given current evidence e^c and historical evidence data D , the Bayesian strategy is adopted to infer the posterior probability of each possible process status, S_j

$$p(S_j | e^c, D) = \frac{p(e^c | S_j, D) p(S_j | D)}{\sum_{F_j} p(e^c | S_j, D) p(S_j | D)} \quad (30)$$

where $p(e^c | S_j, D)$ is the likelihood probability, $p(S_j | D)$ is the prior probability of process status S_j , and $p(S_j | e^c, D)$ is the posterior probability of S_j under current evidence e^c and historical database D . The process status with a large posterior probability is considered to be the probable internal process status.

Then, according to the realization probability of the unavailable monitor's reading and posterior probability of each possible process status S_j under each realization, the outlier probability of incomplete evidence can be calculated by the full probability method as

$$p(S | y_h, H, D) = \sum_{i=1}^{K_h} p(S | e_i, D) p(e_i | y_h, H, D) \quad (31)$$

Overall, the outlier probability estimation for a missing data point in multisampling rates of plant-wide processes can be executed in four phases.

Phase 1: Likelihood probability estimation, which is performed offline:

(1) Calculate $n(e_k | D_c)$ using the complete dataset D_c of certain process statuses.

(2) Use the complete historical data D_c of each process status to calculate the initial value $\theta_k^0 = \frac{n(e_k | D_c)}{N_c}$, which is set as $\Theta^{old} = \begin{bmatrix} \theta_1^0 & \theta_2^0 & \cdots & \theta_K^0 \end{bmatrix}$.

(3) According to Θ^{old} , obtain $n(e_k | D_{ic}, \Theta^{old})$ through the incomplete dataset based on Equation (28).

- (4) Calculate the new likelihood Θ^{new} by using $n(e_k|D_c)$ and $n(e_k|D_{ic}, \Theta^{old})$ based on Equation (29).
 (5) Check whether the terminating conditions are satisfied; if so, record the final likelihood probability. Otherwise, set Θ^{new} as Θ^{old} and repeat steps (3)–(5).

Phase 2: Offline posterior probability estimation

- (1) Based on offline likelihood probability, the posterior probability of each possible process status S_j under each evidence is obtained by Equation (30).

Phase 3: Realization probability estimation, which is performed online.

- (1) For the current observed part y_h of the incomplete sample, calculate $n(e_i|y_h, D_{y_h})$, $N_{D_{y_h}}$, $\alpha(e_i|y_h)$ and A_{y_h} according to Equations (10)–(13).

- (2) Obtain $n(e_i|y_h, H_{y_h})$ and $N_{H_{y_h}}$ according to online moving horizon and Equations (14)–(16).

- (3) Achieve the realization probability $p(e_i|y_h, H_{y_h}, D_{y_h})$ based on Equation (9).

Phase 4: Online full probability estimation.

Using the realization probability of Equation (9) and posterior probability of each possible process status S_j of Equation (30), the outlier probability of an incomplete evidence can be calculated through Equation (31).

The details are also illustrated in Figure 2.

Figure 2. Flowchart of the proposed outlier detection method.

4. Simulation and Application

In this section, the proposed monitoring scheme is applied to a plant-wide TE benchmark process, which is the general test platform of the process monitoring and diagnosis method [24]. First, obtain the history and online test datasets, both of which contain 960 samplings. The outliers in the history dataset are the 120th, 140th, 240th, 250th, 350th, 360th, 460th, 480th, 600th, 720th, and 840th samples; Table 2 shows the outlier reason for these samplings. The outliers in the online test dataset are also the 120th, 140th, 240th, 250th, 350th, 360th, 460th, 480th, 600th, 720th, and 840th samples, with different fault reasons, listed in Table 3.

Table 2. Outliers in the history dataset.

Outlier	Reason	Type	Is Incomplete Data
120	A feed (stream 1)	Pulse change	No
140	Reactor level	Pulse change	Yes
240	D feed (stream 2)	Pulse change	No
250	Reactor temperature	Pulse change	Yes
350	Purge rate (stream 9)	Pulse change	Yes
360	E feed (stream 3)	Pulse change	No
460	Product separator temperature	Pulse change	Yes
480	A and C feed (stream 4)	Pulse change	No
600	Recycle flow (stream 8)	Pulse change	No
720	Reactor feed rate (stream 6)	Pulse change	No
840	Reactor pressure	Pulse change	No

Table 3. Outliers in the online test dataset.

Outlier	Reason	Type	Is Incomplete Data
120	Product separator level	Pulse change	No
140	Stripper steam flow	Pulse change	Yes
240	Product separator pressure	Pulse change	No
250	Compress work	Pulse change	Yes
350	Reactor cooling water outlet temp	Pulse change	Yes
360	Product separator underflow	Pulse change	No
460	Separator cooling water outlet temp	Pulse change	Yes
480	Stripper level	Pulse change	No
600	Stripper pressure	Pulse change	No
720	Stripper underflow (stream 11)	Pulse change	No
840	Stripper temperature	Pulse change	No

Next, suppose that this process is a multisampling rates system, in which some variables are measured with period T , some with $2T$, and the others with $3T$. Then, put the variables with the same sampling period in the same sub-block so that three sub-blocks in total are established. Perform PCA for each sub-block; thus, related three-dimensional evidence with missing data are generated. Among the history dataset, there are five incomplete samples in every six samples, and the 140th, 250th, 350th, and 460th samples are incomplete data.

After obtaining the history dataset and online test dataset, perform the EM-based likelihood probability estimation, and marginalization-based realization estimation, as well as the Bayesian and full probability-based outlier detection. Figure 3 is the iteration process for the normal status and the outlier status, which shows that convergence is fast. The value of the likelihood for each generation and the final convergence value are shown in Tables 4 and 5 for normal status and outlier status, respectively. For normal status, only e_1, e_3, e_4 , and e_6 appear, and the likelihood probability of e_1 is approximately 100%. For outlier status, it is possible for all evidence, except e_1 , to appear with different probabilities.

Figure 3. EM iteration results for normal and outlier status: (a) the normal status; (b) the outlier status.

Table 4. EM iteration results for normal status.

	E1	E2	E3	E4	E5	E6	E7	E8
1	0.993464	0	0	0	0	0.006536	0	0
2	0.998419	0	0.000263	0.000263	0	0.001054	0	0
3	0.998309	0	0.00033	0.000308	0	0.001054	0	0
4	0.998284	0	0.000356	0.000306	0	0.001054	0	0
5	0.998278	0	0.000373	0.000295	0	0.001054	0	0
6	0.998275	0	0.000389	0.000282	0	0.001054	0	0
7	0.998274	0	0.000403	0.000269	0	0.001054	0	0

Table 5. EM iteration results for outlier status.

	E1	E2	E3	E4	E5	E6	E7	E8
1	0	0.7142	0	0.1429	0	0.1429	0	0
2	0	0.3896	0.0714	0.1169	0.0714	0.2078	0.0714	0.0714
3	0	0.3896	0.0714	0.1169	0.0949	0.1845	0.0714	0.0714
4	0	0.3896	0.0714	0.1169	0.1026	0.1770	0.0714	0.0714
5	0	0.3896	0.0714	0.1169	0.1047	0.1745	0.0714	0.0714
6	0	0.3896	0.0714	0.1169	0.1055	0.1737	0.0714	0.0714
7	0	0.3896	0.0714	0.1169	0.1058	0.1734	0.0714	0.0714

The method proposed in this paper is compared with the traditional Bayesian method which uses the complete data only. The comparison contains two aspects:

- (1) The first aspect is the detection result for incomplete samples. The Bayesian detection method, which is based only on complete data, cannot achieve detection, whereas the method adopted in this study can realize the detection for the 140th, 250th, and 350th points, but still deems the 460th point normal instead of classifying it as an outlier, as it is shown in Table 6.
- (2) The second aspect is that, for complete outlier points, both the traditional Bayesian method and the method adopted in this study can achieve detection, but the detection result is slightly different. For the 120th and 480th points, both methods fail to detect the abnormality, while for the 240th and 600th points, the method used in this study finds the fault with higher probability. For the other outlier points, the two methods achieve the same result, which is illustrated in Figure 4.

Table 6. Outlier detection result for incomplete sampling.

Sampling Point	140	250	350	460
Probability of Outlier	1	1	0.933	0.006

Figure 4. Outlier detection result for complete outlier sampling.

Figure 5 is the result for online outlier detection using the adopted method. It reveals that, for all normal points, the correct classifications are given. For outliers in the test dataset, there are two complete outlier points (out of seven outliers), and one incomplete outlier point (out of four points) are wrongly considered a normal point, respectively, so the detection rates are 71.43% and 75% for complete outliers and incomplete outliers, respectively, giving an overall detection rate of 72.8%.

Figure 5. Online detection result for outlier sampling.

5. Conclusions

For effective process control and analysis, it is necessary to eliminate outliers. However, due to the complexity of plant-wide processes, systems usually feature multiple sampling rates, while the traditional outlier detection method typically assumes that the sampling time and interval are the same. Thus, dealing with the problem of multisampling rates is a key issue. In this study, a missing data probability estimation-based Bayesian outlier detection method is proposed for outlier detection in a process with multisampling rates.

The research includes four aspects: (1) If all variables are used as a source of evidence for the Bayesian outlier detection system, the amount of evidence is huge. In order to simplify the calculation while achieving desired rates of detection, variables with the same sampling period are placed in the same sub-block, PCA is performed for each sub-block, and the related T^2 and SPE are used to form monitoring evidence. (2) Due to the different sampling periods, the evidence formed through PCA features missing values, so marginalization-based probability estimation is adopted to calculate the realization of incomplete evidence through evidence from historical multisampling rates and an online moving horizon. (3) The possible internal status for each realization is estimated through the EM algorithm with missing data. (4) Via Bayesian and full probability theories, based on the results of (2) and (3), it is determined whether the current incomplete evidence is an outlier.

The difficulty of this work lies in the means of estimating the most likely realization of incomplete evidence based on historical incomplete evidence and online moving incomplete evidence, for which the EM method for multiple missing data patterns is needed. The main work of this research is thus the detection of outliers for incomplete evidence, using Bayesian and full probability theories, based on the possible realization of incomplete evidence in a multisampling rates system. Nonetheless, some disadvantages of this method remain. The first is that considering the accuracy of the estimation using the EM algorithm, there is a limit to the ratio of the maximum incomplete data that is allowed. Second, since the dimension of evidence is the same as the number of sub-blocks of the system, the sub-block is divided according to the sampling cycles. Thus, too many different sampling cycles leads to higher dimensions of evidence, resulting in the number of possible realizations increasing exponentially, thereby challenging computer capabilities. As a result, there is also a limit to the amount of sampling cycles.

Author Contributions: Conceptualization, Y.T.; Methodology, Y.T.; Software, Y.T., Z.Y., and M.H.; Validation, Y.T., Z.Y., and M.H.; Data curation, Y.T. and Z.Y.; Writing—original draft preparation, Y.T.

Funding: This work was sponsored by the Shanghai Sailing Program (No. 17YF1428300, 17YF1413100) and the Shanghai University Youth Teacher Training Program (ZZslg16009).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Knorr, E.M.; Ng, R.T. Algorithms for Mining Distance-based Outliers in Large Datasets. In Proceedings of the International Conference on very Large Data Bases, New York, NY, USA, 24–27 August 1998; pp. 392–403.
2. Xue, Z.; Shang, Y.; Feng, A. Semi-supervised outlier detection based on fuzzy rough C-means clustering. *Math. Comput. Simul.* **2010**, *80*, 1911–1921. [[CrossRef](#)]
3. Englund, C.; Verikas, A. A hybrid approach to outlier detection in the offset lithographic printing process. *Eng. Appl. Artif. Intell.* **2005**, *18*, 759–768. [[CrossRef](#)]
4. Han, S.J.; Cho, S.B. Evolutionary neural networks for anomaly detection based on the behavior of a program. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **2006**, *36*, 559–570.
5. Hung, W.L.; Yang, M.S. An omission approach for detecting outliers in fuzzy regression models. *Fuzzy Sets Syst.* **2006**, *157*, 3109–3122. [[CrossRef](#)]
6. Lin, C.C.; Chen, A.P. Fuzzy discriminant analysis with outlier detection by genetic algorithm. *Comput. Oper. Res.* **2004**, *31*, 877–888. [[CrossRef](#)]

7. Xu, N.; Zhang, Y. An Efficient Reduction Algorithm of High-dimensional Decision Tables Based on Rough Sets Theory. In Proceedings of the Intelligent Control and Automation (WCICA 2004), Hangzhou, China; 15–19 June 2004; Volume 4305, pp. 4304–4308.
8. Li, X.; Rao, F. An rough entropy based approach to outlier detection. *J. Comput. Inf. Syst.* **2012**, *8*, 10501–10508.
9. Zhang, Y.; Meratnia, N.; Havinga, P.J.M. *Distributed Online Outlier Detection in Wireless Sensor Networks Using Ellipsoidal Support Vector Machine*; Elsevier Science Publishers B.V.: Amsterdam, The Netherlands, 2013; pp. 1062–1074.
10. Little, R.J.A.; Rubin, D.B. *Statistical Analysis with Missing Data, Second Edition*; John Wiley & Sons: Hoboken, NJ, USA, 2002; pp. 200–220.
11. Grzymalabuse, J.W.; Hu, M. *A Comparison of Several Approaches to Missing Attribute Values in Data Mining*; Springer: Berlin/Heidelberg, Germany, 2000; pp. 378–385.
12. Kumar, N.; Hoque, M.A.; Shahjaman, M.; Islam, S.M.S.; Mollah, M.N.H. A new approach of outlier-robust missing value imputation for metabolomics data analysis. *Curr. Bioinform.* **2017**, *12*, [CrossRef]
13. Kim, I.S.; Jung, W. Method of processing the outliers and missing values of field data to improve RAM analysis accuracy. *J. Appl. Reliab.* **2017**, *17*, 264–271.
14. Xiao, H.; Huang, D.; Pan, Y.; Liu, Y.; Song, K. Fault diagnosis and prognosis of wastewater processes with incomplete data by the auto-associative neural networks and ARMA Model. *Chemometr. Intell. Lab. Syst.* **2016**, *161*, 96–107. [CrossRef]
15. Yan, Y.T.; Zhang, Y.P.; Zhang, Y.W.; Du, X.Q. A selective neural network ensemble classification for incomplete data. *Int. J. Mach. Learn. Cybern.* **2016**, *8*, 1–12. [CrossRef]
16. Nowicki, R. On Combining Neuro-Fuzzy architectures with the rough set theory to solve classification problems with incomplete data. *IEEE Trans. Knowl. Data Eng.* **2008**, *20*, 1239–1253. [CrossRef]
17. Luo, C.; Li, T.; Yao, Y. Dynamic probabilistic rough sets with incomplete data. *Inf. Sci.* **2017**, *417*, 39–54. [CrossRef]
18. Pernestaal, A. Probabilistic Fault Diagnosis: With Automotive Applications. Ph.D. Thesis, Linköping University, Linköping, Sweden, 2009; pp. 38–49.
19. Huang, B. Bayesian methods for control loop monitoring and diagnosis. *J. Process Control* **2008**, *18*, 829–838. [CrossRef]
20. Qi, F.; Huang, B.; Tamayo, E.C. A Bayesian approach for control loop diagnosis with missing data. *AIChE J.* **2010**, *56*, 179–195. [CrossRef]
21. Zhang, K.; Gonzalez, R.; Huang, B.; Ji, G. An expectation maximization approach to fault diagnosis with missing data. *IEEE Trans. Ind. Electron.* **2015**, *62*, 1231–1240. [CrossRef]
22. Jiang, Q.; Huang, B.; Ding, S.X.; Yan, X. Bayesian fault diagnosis with asynchronous measurements and its application in networked distributed monitoring. *IEEE Trans. Ind. Electron.* **2016**, *63*, 6316–6324. [CrossRef]
23. Ge, Z.; Song, Z. Distributed PCA model for plant-wide process monitoring. *Ind. Eng. Chem. Res.* **2013**, *52*, 1947–1957. [CrossRef]
24. Downs, J.J.; Vogel, E.F. A plant-wide industrial process control problem. *Comput. Chem. Eng.* **1993**, *17*, 245–255. [CrossRef]

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).