Monitoring of Bioreactor using Statistical Techniques

Damarla Seshu Kumar1 and Kundu Madhusree2
1 Department of Chemical Engineering, NIT Rourkela, Orissa-769008, INDIA
2 Department of Chemical Engineering, NIT Rourkela, Orissa-769008, INDIA

Available online at: www.isca.in
(Received 21st April 2011, revised 2nd May 2011, accepted 28th May 2011)

Abstract

Present study addresses the monitoring of a continuous bioreactor operation. New methodologies; based on clustering time series data and moving window based pattern matching have been proposed for the detection of fault in the chosen bioreactor process. A modified k-means clustering algorithm using similarity measure as a convergence criterion has been adopted for discriminating among time series data pertaining to various operating conditions. The proposed distance and PCA based combined similarity along with the moving window approach were used to discriminate among the normal operating conditions as well as detection of fault for the process taken up.

Keywords: PCA, moving window, pattern matching, bioreactor, k-means clustering

Introduction

Monitoring a chemical process is a challenging task because of their multivariate and highly correlated nature. The data based approaches; supervised learning; unsupervised learning and multivariate statistical techniques rather than the model based approaches are convenient for process monitoring. New methodologies; based on clustering time series data and moving window based pattern matching have been proposed for detection of normal as well as faulty conditions in the process. Data collection has become a mature technology over the years but the analysis of process historical database has become an active area of research.1-3 Bioreactor control and monitoring has been an active area of research over a decade or so. For optimization of cell mass growth and product formation, continuous mode of operation of bioreactors are desirable not the traditional fed batch bioreactors. Several researchers have studied the continuous bioreactor problem4-7. For the bioreactor process 26 numbers of datasets were created comprising both normal and abnormal operating conditions. At a given time period of interest; for a multivariate time series data or template data, a similar pattern can be located in the historical database using the proposed pattern matching algorithm. A modified k-means clustering algorithm using similarity measures as a convergence criterion has been used for clustering datasets pertaining to different operating conditions including faulty one. Both the pattern matching and clustering time series data are useful for successful monitoring of the process including fault detection and its analysis.

Theoretical Postulations

Similarity Factors: PCA similarity: Principal component analysis is multivariable statistical technique to reduce the dimensionality of the large datasets by transforming set of original correlated variables into a new set of uncorrelated variables. These new uncorrelated variables capture the maximum variance in the dataset and are linear combinations of the original variables. PCA was successfully applied to cluster multivariate time series data8,9. PCA similarity factor was developed by choosing largest k principal components of each multivariate time series dataset that describe at least 95 % of variance in the each dataset10. These
principal components are the eigen vectors of the 
covariance matrix. The PCA similarity factor 
between two datasets is defined by equation (1) 
\[
S_{PCA} = \frac{1}{\kappa} \sum_{i=1}^{\kappa} \sum_{j=1}^{\kappa} \cos^2 \theta_{ij}
\]

(1)

Where \( \kappa \) is the number of selected principal 
components in both datasets, \( \theta_{ij} \) is the angle 
between the \( i^{th} \) principal component of \( X_1 \) and 
\( j^{th} \) principal component of \( X_2 \). When first two 
principal components explain 95% of variance in 
the datasets, \( S_{PCA} \) may not capture the degree of 
similarity between two datasets because it weights 
all PCs equally. Obviously \( S_{PCA} \) has to modify to 
weight each PC by its explained variance. The 
modified \( S^2_{PCA} \) is defined as 
\[
S^2_{PCA} = \sum_{i=1}^{\kappa} \sum_{j=1}^{\kappa} (\lambda_i^2 \lambda_j^2) \cos^2 \theta_{ij}
\]
\[
\sum_{i=1}^{\kappa} \sum_{j=1}^{\kappa} \lambda_i^2 \lambda_j^2
\]

Where \( \lambda_i^2, \lambda_j^2 \) are the Eigen values of the first and 
second datasets respectively.

**Distance similarity:** In addition to above 
similarity measure, distance similarity factor can 
be used to cluster multivariate time series data. 
Distance similarity factor compares two datasets 
that may have similar spatial orientation. The 
process variables pertaining to different operating 
conditions may have similar principal components.

The distance similarity factor is defined as 
\[
S_{dist} = 2 \times \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz = 2 \times \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2} dz
\]

(3)

Where \( \Phi = (\bar{x}_2 - \bar{x}_1) \Sigma_1^{-1} (\bar{x}_2 - \bar{x}_1)^T \), \( \bar{x}_2 \) & \( \bar{x}_1 \) 
are sample means row vectors.

\( \Sigma_1 \) is the covariance matrix for dataset \( X_1 \) and 
\( \Sigma_1^{-1} \) is pseudo inverse of \( X_1 \). Dataset \( X_1 \) is 
assumed to be a reference dataset. In equation (3), 
a one side Gaussian distribution is used because 
\( \Phi \geq 0 \). The error function can be calculated by 
using any software or standard error function 
tables. The integration in equation (3) normalizes 
\( S_{dist} \) between zero and one.

**Combined similarity Factor:** The combined 
similarity factor \( (SF) \) combines \( S^2_{PCA} \) and \( S_{dist} \) using 
weighted average of the two quantities and used 
for clustering of multivariate time series data. The 
combined similarity is defined as 
\[
SF = \alpha_1 S^2_{PCA} + \alpha_2 S_{dist}
\]

(4)

The selection of \( \alpha_1 \) and \( \alpha_2 \) is up to the user but 
ensure that the sum of them is equal to one. In this 
work we selected the values of \( \alpha_1 \) & \( \alpha_2 \) are 0.67 
and 0.33.

**K-means clustering using similarity factors:**

The time series data pertaining to various operating 
conditions were discriminated and classified using 
the following similarity based K-means algorithm.

Given: Q datasets, \( \{x_{i1},x_{i2},\ldots,x_{iq}\} \) to be clustered 
into k clusters

1. Let \( j^{th} \) dataset in the \( i^{th} \) cluster be defined as \( x_{ij} \). 

Computation of the aggregate dataset \( X_i (i = 1,2,\ldots,k) \), 
for each of the k clusters as, 
\[
X_i = [(x_{i1})^T \ldots (x_{ij})^T \ldots (x_{iq})^T]^T
\]

(5)

Where \( Q_i \) is the number of datasets in the 
database. \( \sum_{i=1}^{k} Q_i = Q \).

2. Calculation of the dissimilarity between dataset 
\( x_q \) and each of the k aggregate datasets 
\( X_i, i = 1,2,\ldots,k \) as, 
\[
d_{i,q} = 1 - SF_{i,q}
\]

(6)

Where \( SF_{i,q} \) is similarity between \( q^{th} \) dataset and 
\( i^{th} \) cluster described by equation 4. Let the 
aggregate dataset \( X_i \) in equation 6 be the reference 
dataset. Dataset \( x_q \) is assigned to the cluster to 
which it is least dissimilar. Repetition of this step 
for Q datasets.

**Clustering performance evaluation:** Some key 
definitions were introduced to evaluate the 
performance of the clusters obtained using 
similarity factors\(^{11}\). Assuming the number of 
operating conditions is \( N_{op} \) and the number of
datasets for operating condition \( j \) in the database is \( N_{DBj} \). Cluster purity is defined to characterize each cluster in terms of how many numbers of datasets for a particular operating condition present in the \( i^{th} \) cluster. Cluster purity is defined as,

\[
P_{j}^{p} = \left( \frac{\Delta N_{ij}}{N_{pi}} \right) \times 100\%
\]

(7)

Where \( N_{ij} \) is the number of datasets of operating condition \( j \) in the \( i^{th} \) cluster and \( N_{pi} \) is the number of datasets in the \( i^{th} \) cluster.

Cluster efficiency measures the extent to which an operating condition is distributed in different clusters. This method is to penalize the large values of \( k \) when operating condition \( j \) distributed into different clusters. Clustering efficiency is defined as,

\[
\eta = \left( \frac{\max N_{ij}}{N_{DBj}} \right) \times 100\%
\]

(8)

Where \( N_{DBj} \) is the number of datasets for operating condition \( j \) in the database. Large number of datasets of operating condition present in a cluster can be considered as dominant operating condition.

Moving Window Based Pattern Matching: In this approach, the snapshot or template data with unknown start and end time of operating condition moves through historical data and the similarity between them is characterized by distance and PCA based combined similarity factor. The historical data windows with the largest values of similarity factors are collected in a candidate pool and are called records to be analyzed by the process Engineer. For the present work, the historical data window moved one observation at a time, with each old observation is getting replaced by new one. Pool accuracy, Pattern matching efficiency and overall effectiveness of pattern matching are important metrics that quantify the performance of the proposed pattern matching algorithm.

\( N_{P} \): The size of the candidate pool. \( N_{P} \) is the number of historical data windows that have been labeled ‘similar’ to the snapshot data by a pattern matching technique. The data windows collected in the candidate pool are called records.

\( N_{i} \): number of records in the candidate pool that are exactly similar to the current snapshot data, i.e. having a similarity of 1.0/or number of correctly identified record.

\( N_{2} \): number of records in the candidate pool that are not correctly identified.

\( N_{P} = N_{i} + N_{2} \)

\( N_{DB} \): The total number of historical data windows that are actually similar to the current snapshot. In general, \( N_{DB} \neq N_{P} \)

Pool accuracy\( = \frac{N_{i}}{N_{P}} \times 100\% \)

Pattern matching efficiency \( = \left[ 1 - \frac{(N_{P} - N_{i})}{N_{DB}} \right] \times 100\% \)

Pattern matching algorithm efficiency \( = \frac{N_{P}}{N_{DB}} \times 100\% \)

A large value of Pool accuracy is important in case of detection of small number of specific previous situations from a small pool of records without evaluating incorrectly identified records. A large value of Pattern matching efficiency is required in case of detection of all of the specific previous situations from a large pool of records. The proposed method is completely data driven and unsupervised; no process models or training data are required. The user should specify only the relevant measured variables.

Model of Bioreactor: A (2×2) bioreactor process was taken up. The primary aim of a continuous bioreactor is to avoid wash out condition which ceases reaction that may be achieved either by controlling cell mass (\( X \) g/L) or substrate concentrations (\( S \) g/L). Dilution rate \( (D=F/V \text{ (h}^{-1}) \)) and feed substrate concentration \( (S_{f}, \text{ g/L}) \) are served as manipulated variables to control cell mass \( (X \text{ g/L}) \) or substrate concentrations \( (S \text{ g/L}) \). Thus two degrees of freedom is available for control. The study is based on single biomass-single substrate process. The following are the model equation based on first principle.

\[
\frac{dx_{1}}{dt} = (\mu - D)x_{1}
\]

(9)

\[
\frac{dx_{2}}{dt} = D(x_{2f} - x_{2}) - \frac{\mu x_{1}}{Y}
\]

(10)
The reaction rate is given by $\eta = \mu x_1$ \hspace{1cm} (11)

Where $x_{1f}$ is the substrate concentration in the feed. $x_1$ & $x_2$ are the biomass and substrate composition, respectively. $\mu$, the specific growth is a function of substrate concentration and given by the substrate inhibition growth rate expression:

$$\mu = \frac{\mu_{\text{max}} x_2}{k_m + x_2 + k_i x_2^2}$$ \hspace{1cm} (12)

The relation between the rate of generation of cells and consumption of nutrients is defined by the yield given in the following equation

$$\gamma = \frac{r_1}{r_2}$$ \hspace{1cm} (13)

Introducing the dilution rate ($D = \frac{F}{V}$) and assuming there is no biomass in the feed, i.e., $x_{1f} = 0$.

The inputs are dilution rate and feed substrate concentration and the outputs are the concentrations of substrate and biomass (All values in deviation form). The values of steady state dilution rate ($D_s$), feed substrate concentration ($x_{2f}$), the steady state values of the states at the stable and unstable operating points and the various parameters are presented in table 1. When both the concentrations (biomass & substrate) are high process leads to unstable equilibrium. When there is substrate limiting condition, process is at stable equilibrium. Direct synthesis controllers were designed to control the biomass and substrate concentration in both stable and unstable situations. Open loop and closed loop processes were considered in order to generate the database including faults using distinct operating conditions at stable & unstable operating points (by varying the controller tuning parameter, $\lambda$, the faulty operating condition 4 was generated). Bioreactor was simulated for one hour and data was taken up with a sampling interval of 6 seconds using different operating conditions. Total database contains 26 datasets of various operating conditions where each dataset contains 600 observations of two outputs and are presented in table 2.

**Results and Discussion**

Table 2 presents the 26 numbers of datasets, which were generated for various operating conditions by varying the parameters $k_m$, $k_i$ & $\lambda$. $\lambda$ is the tuning parameter of direct synthesis controller, $k_m$ - a parameter (both for Monod and substrate inhibition), $k_i$ - a parameter for substrate inhibition only. Four numbers of optimum clusters were obtained using similarity based; modified $K$-means algorithm. Faulty operating condition 4 was well captured by cluster 4. The derived cluster purity and efficiency; both were 100 % as presented in table 3.

Pattern matching was done using the moving window in a sample wise manner. 4 sets of database pertaining to four various operating conditions were considered as historical database and 4 snapshot data sets were considered. Pool accuracy and Pattern matching efficiency were determined to be 100 %. Similarity factors in the range of 0.965 to 1.0 were considered in this work. Table 4 presents the proposed pattern matching performance.

**Conclusions**

A moving window based pattern matching technique was developed with a view to process monitoring. The proposed approach successfully located the arbitrarily chosen different operating conditions of current period of interest among the historical database of a continuous bioreactor process. The PCA and distance based similarity factors provided the effective way of pattern matching in a multivariate time series database of a bioreactor process. The time series data pertaining to various operating conditions of the considered bioreactor process were discriminated /classified efficiently using a similarity based modified $K$-means clustering algorithm. The present developments can be considered as effective data based tool/machine learning algorithm for process monitoring.

**Acknowledgements**

Authors would like to thank to Prof. Palash Kundu, EE Dept., Jadavpur University, Kolkata, India for providing motivation for this work.
References


Table-1: Bioreactor process Parameters

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_{max}$</td>
<td>0.53 h$^{-1}$</td>
</tr>
<tr>
<td>$k_m$</td>
<td>0.12 g/L</td>
</tr>
<tr>
<td>$k_i$</td>
<td>0.4545 L/g</td>
</tr>
<tr>
<td>$y$</td>
<td>0.4</td>
</tr>
<tr>
<td>$x_{2fs}$</td>
<td>4.0 g/L</td>
</tr>
<tr>
<td>$D_i$</td>
<td>0.3 h$^{-1}$</td>
</tr>
<tr>
<td>$x_{1s}$ (at stable operating point)</td>
<td>1.5302 g/L</td>
</tr>
<tr>
<td>$x_{2s}$ (at stable operating point)</td>
<td>0.1746 g/L</td>
</tr>
<tr>
<td>$x_{1u}$ (at unstable operating point)</td>
<td>0.995103 g/L</td>
</tr>
<tr>
<td>$x_{2u}$ (at unstable operating point)</td>
<td>1.512243 g/L</td>
</tr>
</tbody>
</table>
Table 2: Database corresponding to various operating conditions for Bio-Chemical Reactor process

<table>
<thead>
<tr>
<th>Op. Cond.</th>
<th>Parameter range</th>
<th>No. of datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$0.04545 \leq K \leq 0.4545$</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>$0.015 \leq K_m \leq 0.12$</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>$0.2208 \leq \lambda \leq 1.1043$</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>$2.9446 \leq \hat{\lambda} \leq 8.83838$</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 3: Combined similarity factor based clustering performance

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>100</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>100</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3 (faulty cond.)</td>
</tr>
</tbody>
</table>

Avg. $P = 100$, $\eta = 100$, $\eta = 100$, $\eta = 100$,

Table 4: Moving window based pattern matching performance

<table>
<thead>
<tr>
<th>Snapshot Op. Cond.</th>
<th>Size of the candidate pool $N_p$</th>
<th>$N_1$</th>
<th>$N_2$</th>
<th>Pattern matching efficiency</th>
<th>Pool Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Op. Cond. 1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Op. Cond. 2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Op. Cond. 3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Op. Cond. 4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>