Pattern Matching in Historical Data

Michael C. Johannesmeyer
Procter & Gamble Manufacturing Company, Sacramento, CA 95826

Ashish Singhal and Dale E. Seborg
Dept. of Chemical Engineering, University of California, Santa Barbara, CA 93106

For many engineering and business problems, it would be very useful to have a general strategy for pattern matching in large databases. For example, the analysis of an abnormal plant condition would benefit if previous occurrences of the abnormal condition could be located in the historical data. A new pattern-matching strategy is proposed for multivariate time series based on statistical techniques, especially principal-component analysis (PCA). The new approach is both data-driven and unsupervised because neither training data nor a process model is required. Given an arbitrary set of multivariate data, the new approach can be used to locate similar patterns in a large historical database. The proposed pattern-matching strategy is based on two similarity factors: the standard PCA similarity factor and a new similarity factor that characterizes the pattern of alarm violations. An extensive simulation study for a chemical reactor demonstrates that this strategy is more effective than existing PCA methods and can successfully distinguish between 28 different operating conditions.

Introduction

Due to significant advances in data collection and storage, enormous amounts of data are routinely collected and stored in fields such as engineering, business, and bioinformatics. In large industrial plants, sensor data are recorded for thousands of process variables as often as every second. Product quality, production, and maintenance data are stored less frequently, often in separate databases. Thus, massive amounts of data are available for analysis. Historical databases contains potentially valuable process information, but it is notoriously difficult to extract. Because industrial plants are "data rich, but information poor," an important research question is: "How can relevant information be extracted from vast amounts of historical data in order to better understand processes and to monitor their performance?" Similar questions are receiving considerable attention in other fields, as indicated by the growing interest in data mining and knowledge discovery problems (Apté, 1997; Kennedy et al., 1998; Ramakrishnan et al., 2000).

This article is concerned with the following general pattern-matching problem. Given an arbitrary set of multivariate time-series data, how can similar patterns be located in a large database? Although this type of problem can arise in many different engineering and business situations, our focus is on manufacturing data from industrial processes. One important application involves process monitoring and the diagnosis of abnormal situations. If the same type of abnormal plant operation has occurred in the past, then the relevant historical data provide a valuable source of information. The ability to locate data records for previous abnormal situations will be advantageous for difficult process diagnosis problems. This additional information can facilitate two important activities: (1) identifying the root cause of the abnormal operation, and (2) developing an effective remedy that will prevent future abnormal situations or at least minimize their impact.

In this article a novel pattern-matching methodology is developed for multivariate time-series data. The proposed methodology provides a preliminary screening of historical data in order to locate previous periods of similar, but not necessarily identical, process behavior. Neither a process model nor training data are required for pattern matching. In particular, the new approach is not a "supervised learning technique," unlike most neural-network and pattern recognition methods. Instead, the proposed pattern-matching strategy relies on novel data analysis and pattern-recognition techniques to analyze historical data in an efficient manner while requiring only a modest amount of computer resources.

Correspondence concerning this article should be addressed to D. E. Seborg.
Current address of A. Singhal: Johnson Controls Inc., 507 E. Michigan St., Mail Box M36, Milwaukee, WI 53202.

Motivation

Plant personnel have long recognized the value of their plant data and have invested significant resources to establish large historical databases. But despite industrial interest and significant potential benefits, the information contained in these databases has remained very difficult to extract. The key issue continues to be: How can relevant information be located in such a vast sea of data? In order to locate relevant data in such large databases, techniques must be developed that require modest computational effort while still revealing the unique and relevant characteristics of the data.

Sometimes it is relatively easy to search historical data in order to locate previous occurrences of an abnormal plant operation. For example, the incident might be associated with specific operating conditions such as a particular type of grade change, catalyst, or raw material. Similarly, efficient searches could be conducted if the abnormal plant operation can be characterized by simple criteria such as unusual alarm violations. But other types of abnormal situations are not so easily characterized, and plant personnel may have only vague recollections of when previous incidents occurred (e.g., "These types of sustained control loop oscillations have occurred a couple of times during the past two years"). Furthermore, some previous incidents may not have been detected at all.

If historical data are searched manually, the search tends to be tedious and very time-consuming unless it is restricted to a relatively small time period, for example, a few hours or days, instead of months or years. In view of the huge databases and small numbers of process engineers in most plants, it is unlikely that someone familiar with the process would have the time to conduct a manual search. Thus, manual searches of large databases will only be performed under compelling circumstances, for example, after a plant accident or a serious product quality problem. These considerations motivate the present research.

The objective of this article is to develop an effective pattern-matching strategy that can be used to provide a preliminary screening of a large database. Emphasis is placed on a preliminary screening of historical data for the following reason. If an efficient screening technique were available, it could be used to narrow the search for similar periods of process behavior by identifying a relatively small number of promising data records within the historical database. These records will be referred to as the *candidate pool*. Then a person familiar with the process (a *process expert*) could evaluate the candidate pool in order to discern patterns and to diagnose the root cause of the abnormal operation. For example, the process expert might be able to quickly eliminate some records in the candidate pool based on knowledge of the plant history (e.g., equipment problems, feedstock changes) and access to information that is not contained in the historical database, such as operator logs and plant maintenance records.

Ideally, the new methodology should be perfect in the sense that it locates all previous periods of abnormal behavior without any "false positives." But a more realistic objective is to develop a new methodology that provides a preliminary screening of the historical data in order to generate a small candidate pool for subsequent analysis by a person who is familiar with the process. This objective is the primary goal for this research.

An analogous approach is used by law enforcement agencies for fingerprint matching. Typically, a preliminary database search is performed electronically, but manual inspection of the search results is required before a positive identification is confirmed (Anonymous, 2000).

Previous Work

Pattern matching in historical data can be viewed in the broader context of data mining and pattern recognition. In recent years, the terms *data mining* and *knowledge discovery* have become buzzwords for a variety of activities, conferences, and commercial products that pertain to the analysis of large databases (Ramakrishnan et al., 2000). The primary goal of data mining is to discover previous unknown relationships in databases (Fayyad et al., 1996), or to learn from similar patterns in the historical databases (Wang, 2001). Data-mining techniques include classification algorithms, clustering techniques, decision tree methods, rule-based systems, and neural networks (Shürmann, 1996).

Recent data-mining conferences and journal articles have reported business applications involving customer data, sales records, and transaction histories. Data-mining applications have been reported in other fields such as aerospace, seismic data interpretation, and stock market analysis (Fayyad et al., 1996; Apté, 1997; Gavrilov et al., 2000; Ramakrishnan et al., 2000). However, engineering applications are conspicuously absent from data-mining publications. Finally, there is intense current interest in using data-mining techniques to discover the relationships between chemical structures and biological functions in molecular biology and medicine (Dehaspe et al., 1998; Regaldo, 1999).

A number of techniques have been developed for pattern matching in time-series data (Faloutsos et al., 1994; Agrawal et al., 1995; Keogh et al., 2001), including stock market data (Gavrilov et al., 2000). Wang and McGreavy (1998) have clustered multivariate time series using the Expectation Maximization (EM) algorithm (Dempster et al., 1977). Their objective was to cluster similar periods of operation in a historical database for further analysis. Smyth (1999) has reported a similar probabilistic clustering of time-series data using the EM algorithm that evaluates static and dynamic features of the multivariate data. Perng et al. (2000) have developed a "Landmarks" similarity model that calculates the similarity of time-series data using transformations of extracted Landmarks features. However, the methodologies of Smyth (1999) and Perng et al. (2000) are restricted to univariate time-series data.

The subject of the present article can be considered to be a problem in pattern recognition, or more specifically, in *pattern matching*. There is an extensive literature on pattern-recognition techniques and their applications in areas such as machine learning, image processing, and speech and character recognition (Fukunaga, 1990; Bishop, 1995; Shürmann, 1996). Survey articles on process monitoring and data analysis (Kramer and Mah, 1994; Stephanopoulos and Han, 1994; Davis et al., 1999) have considered pattern-recognition techniques from a chemical engineering perspective. An industrial consortium for abnormal situation management (ASM) has evaluated a wide variety of strategies for process monitoring and fault diagnosis (Mylaraswamy, 2001). A key issue in pattern recognition is whether training data are available to facilitate supervised learning. Most of the standard pattern-recognition methods are based on supervised learning.

If training data are available, classification techniques can be devised using powerful techniques such as neural networks, statistical approaches, and rule-based systems (Duda and Hart, 1973; Fukunaga, 1990; Bishop, 1995; Shürmann, 1996; Duda et al., 2001). For example, chemical engineering applications of neural networks for fault classification have been reported by Kavuri and Venkatasubramanian (1993) and Sorsa and Koivo (1993). Vedam and Venkatasubramanian (1999) used signed directed graphs (SDG) and principal-component analysis (PCA) for diagnosis of multiple faults. This methodology requires extensive knowledge about the interrelationships between different process variables using either model equations or knowledge from plant experts.

But for the general problem of pattern matching considered in this article, it would be unduly restrictive to assume that previous instances of the current abnormal operation are available to serve as training data. For this reason, the methodology developed in this article is not based on supervised learning. Instead, multivariate statistical techniques such as PCA are utilized due to the highly correlated nature of plant data. The survey article by Davis et al. (1999) provides a comprehensive review of both supervised as well as unsupervised methodologies for data analysis, including statistical techniques, neural networks, and knowledge-based systems.

For the present research, the closest literature references are the articles by Raich and Çinar (1994, 1995, 1996, 1997). They developed innovative methods for discriminating between different types of faults using standard PCA metrics and the PCA similarity factor. Their approach relies on building PCA models using representative data for each type of disturbance or fault that is to be considered. Our methodology differs from that of Raich and Çinar (1996, 1997), because we are concerned with pattern matching rather than fault diagnosis. Consequently, we do not assume that known fault patterns are available *a priori*. Also, they calculate PCA metrics for each sample in the current data set, while our methodology is based on matching two sets of data and characterizing their degree of similarity. Furthermore, our proposed pattern-matching strategy requires neither training data nor *a priori* knowledge.

Certain types of neural-network techniques, such as adaptive resonance theory (ART) and self-organizing maps, can be used for unsupervised learning (Koivo, 1994; Whiteley and Davis, 1994). But these methods require considerable computational effort for large problems, in comparison with multivariate statistical techniques like PCA.

Proposed Approach

The overall pattern-matching strategy is illustrated by the flow chart in Figure 1. First, the user defines the *snapshot* or *template* data that characterize the abnormal plant operation. This involves (1) specifying the relevant process variables, and (2) the time period of interest, that is, the duration of the abnormal operation. Specification of the process variables and the time period defines the snapshot data that are used to locate similar patterns in the historical data.

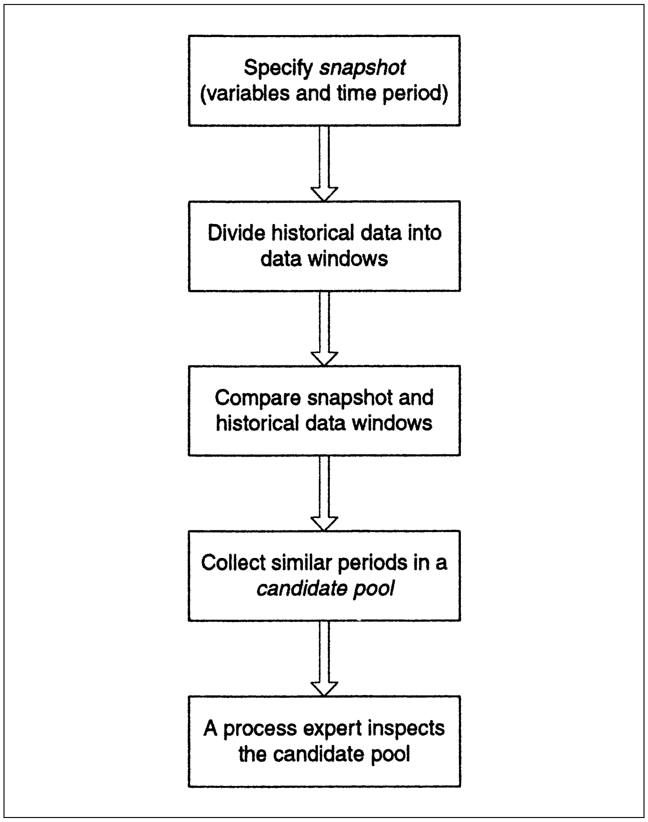
Specify *snapshot* (variables and time period)
Divide historical data into data windows
Compare snapshot and historical data windows
Collect similar periods in a *candidate pool*
A process expert inspects the candidate pool

Figure 1. Proposed data mining approach.

Next, the relevant portion of the historical database is divided into data windows that are the same size as the snapshot data. Thus, the data windows contain the same process variables and the same number of samples as the snapshot data. For this article, the historical data were divided into adjacent, nonoverlapping data windows. Extensions are currently being developed for the more general cases where the historical data are not divided into data sets, and for situations where the start and end times of the operating conditions are not known *a priori* (Singhal, 2002).

After the historical data have been divided into data windows, the snapshot data are compared to each data window using an appropriate *similarity measure*. In this article, similarity factors based on PCA and the number of alarm limit violations are developed and used to quantify the similarity between the snapshot data and historical data windows. The historical data windows that have large similarity factors are labeled as "similar" to the snapshot data.

The historical data windows that are designated as similar to the snapshot data are collected in a *candidate pool*. The records in the candidate pool then can be evaluated by a process expert to gain further insight into plant operation.

Process monitoring based on PCA

PCA is a multivariate statistical technique that calculates the principal directions of variability in data, and transforms a set of correlated variables into a new set of uncorrelated variables (Jolliffe, 1986; Jackson, 1991). The new uncorrelated variables are linear combinations of the original variables. PCA is widely used for monitoring of multivariable processes. The survey papers by Davis et al. (1999), Kourti and MacGregor (1996), and Zhang et al. (1995), and books by Jackson (1991), Wang (1999), and Chiang et al. (2001) describe the PCA methodology and its applications to a wide variety of processes.

Consider an $m \times n$ data matrix, $\boldsymbol{X}$, for n process variables with m measurements of each variable. Assume that the data for each variable have been scaled to zero mean and unit variance. Data matrix $\boldsymbol{X}$ can be expanded using principal components $\boldsymbol{p}_i$, score vectors $\boldsymbol{t}_i$, and a residual matrix, $\boldsymbol{E}$

$$\boldsymbol{X} = \boldsymbol{t}_1 \boldsymbol{p}_1^T + \boldsymbol{t}_2 \boldsymbol{p}_2^T + \cdots + \boldsymbol{t}_k \boldsymbol{p}_k^T + \boldsymbol{E}. \tag{1}$$

Each of the k principal components is an eigenvector of $\boldsymbol{\Sigma}$, the covariance matrix of $\boldsymbol{X}$,

$$\Sigma \boldsymbol{p}_i = \lambda_i \boldsymbol{p}_i \qquad i = 1, 2, \ldots, k, \tag{2}$$

where λ_i is the ith eigenvalue of the covariance matrix

$$\Sigma = \frac{\boldsymbol{X}^T \boldsymbol{X}}{m-1} \tag{3}$$

The principal components are numbered so that $\boldsymbol{p}_1$ corresponds to the largest eigenvalue and $\boldsymbol{p}_k$, the smallest. The variability in the data matrix can be related to the principal components because the variability explained by a principal component is proportional to its eigenvalue (Jackson, 1991). A variety of methods are available for choosing k, the number of principal components (Jackson, 1991; Valle et al., 1999).

The first step in PCA process monitoring is the development of a PCA model using data that corresponds to "normal operation." This model can then be used for subsequent monitoring in the following manner. A new measurement at time j, $\boldsymbol{x}(j)$, is projected onto the PCA model to give the corresponding $1 \times k$ score vector, $\boldsymbol{t}(j)$

$$\boldsymbol{t}(j) = \boldsymbol{x}(j)\boldsymbol{P} \tag{4}$$

The elements of $\boldsymbol{t}(j)$ and the principal components can then be used to provide an estimate of the current data point,

$$\hat{\boldsymbol{x}}(j) = \boldsymbol{t}(j)\boldsymbol{P}^T \tag{5}$$

The difference between this estimate and the actual measurement vector, $\boldsymbol{x}(j)$, is the PCA model error or residual, $\boldsymbol{e}(j)$:

$$\boldsymbol{e}(j) = \boldsymbol{x}(j) - \hat{\boldsymbol{x}}(j) \tag{6}$$

Process monitoring based on PCA typically involves calculation of two statistics, the Q statistic and Hotelling's T^2 statistic (Kourti et al., 1996; Martin and Morris, 1996). The Q statistic at time j is calculated from the current residual:

$$Q(j) = \boldsymbol{e}(j)\boldsymbol{e}(j)^T \tag{7}$$

It provides a measure of how well the new measurement is described by the PCA model. Hotelling's T^2 statistic can be calculated from,

$$T^2(j) = \boldsymbol{t}(j)\boldsymbol{\Lambda}^{-1}\boldsymbol{t}^T(j), \tag{8}$$

where $\boldsymbol{\Lambda}$ is the diagonal eigenvalue matrix with elements, λ_i, defined in Eq. 2. The T^2 statistic provides a measure of the variation within the PCA model for each measurement, $\boldsymbol{x}(j)$. Confidence limits for the Q and T^2 statistics (Wise and Gallagher, 2000) can be calculated based on the assumption that the residuals in Eqs. 1 and 6 are independent and identically distributed (IID).

PCA-based similarity factors

Similarity factors provide measures of the similarity between two data sets. Consider a historical data set $\boldsymbol{H}$ and a current snapshot data set $\boldsymbol{S}$. A PCA model is built for the snapshot dataset $\boldsymbol{S}$ and the 95% confidence limits for the T^2 and Q statistics are calculated. These chart limits are denoted as T_{95}^2 and Q_{95}, respectively. The data for the historical data set $\boldsymbol{H}$ are projected onto the PCA model for the snapshot data, and the T^2 and Q values are calculated for each of the m samples in $\boldsymbol{H}$. The T^2 and Q similarity factors are defined based on the 95% chart limit violations of these statistics. If the number of violations for data set $\boldsymbol{H}$ is less than or equal to a critical number r_{95}, then data sets $\boldsymbol{H}$ and $\boldsymbol{S}$ are considered to be similar. The critical number of limit violations is calculated in the following manner (Singhal and Seborg, 2000).

The numbers of Q and T^2 limit violations for a data window of size m are assumed to follow a binomial distribution with probability parameter, $\epsilon = 0.05$. This premise is consistent with the IID assumption for the data used to build the PCA models. Parameter ϵ is the probability that a single data point violates the 95% chart limits. By definition, $\epsilon = 0.05$ if the limits on the T^2 statistic are 95% limits. The maximum number of T^2 limit violations allowed at a 95% confidence level is denoted by r_{95} and is referred to as the *critical number of violations*. It can be calculated as the inverse of the cumulative binomial distribution at the probability value of 0.95 (Singhal and Seborg, 2000)

$$r_{95} = \max \quad r,$$

such that

$$\sum_{k=0}^{r} \binom{m}{k} \epsilon^k (1-\epsilon)^{m-k} < 0.95. \tag{9}$$

The similarity factors are defined as follows.

(1) *T^2 similarity factor:* Data set $\boldsymbol{H}$ is considered to be similar to data set $\boldsymbol{S}$ if the number of times that the T^2 statistic exceeds T_{95}^2 is less than the critical number of violations, r_{95}. For a data window of $m = 1024$ observations, $r_{95} = 63$.

(2) *Q similarity factor:* Data set $\boldsymbol{H}$ is considered to be similar to data set $\boldsymbol{S}$ if the number of times the Q statistic ex-

ceeds the Q_{95} value is less than the critical number, r_{95}. The critical number of violations is calculated in the same manner as for the T^2 similarity factor. Thus, $r_{95} = 63$ for 1024 observations.

(3) *Combined discriminant similarity factor:* This similarity factor combines the information contained in the T^2 and Q statistics to provide discrimination between fault types (Raich and Çinar, 1994). At the 95% confidence level, the combined discriminant, C, is calculated as

$$C = \beta \left(\frac{Q}{Q_{95}} \right) + (1 - \beta) \left(\frac{T^2}{T_{95}^2} \right), \tag{10}$$

where β is a weighting factor between zero and one. In the absence of any additional information, the T^2 and Q statistics are weighted equally ($\beta = 0.5$). The cutoff for the C statistic is therefore one. Data set $\boldsymbol{H}$ is considered similar to $\boldsymbol{S}$ if the number of times that the C statistic is greater than unity is less than r_{95}. The critical number of violations is calculated in the same manner as for the T^2 similarity factor, and thus, $r_{95} = 63$.

Krzanowski (1979) proposed a *PCA similarity factor*, S_{PCA}, as a measure of the similarity between two data sets. Assume that the two data sets contain the same n variables and that each of the corresponding PCA models has k principal components, where $k \leq n$. The similarity between the two data sets is then quantified by comparing the k principal components for each data set. The appeal of this approach is that the similarity between two data sets is quantified with a single number, S_{PCA}.

Consider a current snapshot data set $\boldsymbol{S}$ and a historical data set $\boldsymbol{H}$ with each data set consisting of m measurements of the same n variables. Let k_1 be the number of principal components that describe at least 95% of the variance in data set $\boldsymbol{S}$, and k_2 be the number of principal components that describe at least 95% of the variance in data set $\boldsymbol{H}$. Let $k = \max(k_1, k_2)$, which ensures that k principal components describe at least 95% of the variance in each data set. Then subspaces of the $\boldsymbol{S}$ and $\boldsymbol{H}$ data sets can be constructed by selecting only the first k principal components for each data set. The corresponding $(n \times k)$ principal component subspaces are denoted by $\boldsymbol{L}$ and $\boldsymbol{M}$, respectively. The matrices $\boldsymbol{L}$ and $\boldsymbol{M}$ are also the eigenvector matrices corresponding to the first k eigenvalues of the covariance matrices of $\boldsymbol{S}$ and $\boldsymbol{H}$, respectively. The PCA similarity factor compares these reduced subspaces and can be calculated from the angles between principal components (Krzanowski, 1979)

$$S_{PCA} = \frac{1}{k} \sum_{i=1}^{k} \sum_{j=1}^{k} \cos^2 \theta_{ij} \tag{11}$$

where θ_{ij} is the angle between the ith principal component of data set $\boldsymbol{S}$ and the jth principal component of data set $\boldsymbol{H}$. It can also be expressed in terms of the subspaces $\boldsymbol{L}$ and $\boldsymbol{M}$ as (Krzanowski, 1979):

$$S_{PCA} = \frac{\text{trace}(\boldsymbol{L}^T \boldsymbol{M} \boldsymbol{M}^T \boldsymbol{L})}{k}. \tag{12}$$

Because $\boldsymbol{L}$ and $\boldsymbol{M}$ contain the k most significant principal components for $\boldsymbol{S}$ and $\boldsymbol{H}$, S_{PCA} is also a measure of the similarity between data sets $\boldsymbol{S}$ and $\boldsymbol{H}$. The two data sets $\boldsymbol{S}$ and $\boldsymbol{H}$ are considered to be similar, if the value of S_{PCA} exceeds a specified cutoff or threshold value.

Alarm limit violation similarity factor

In order to provide further discrimination between data sets, another source of valuable information is utilized, namely, the number of times that a process variable exceeds specified limits (e.g., alarm limits). A *limit violation similarity factor*, S_{LV}, is proposed to quantify the similarity between two data sets by comparing which variables exceed specified limits and which ones do not. Distributed control systems (DCS) allow high and low alarm limits to be specified for each measured or calculated variable. Alarm limit violations can then be stored in a data historian. However, the specified limits for pattern matching can be selected using other criteria, such as three-sigma limits based on past operating experience.

For the purposes of this study, 95% confidence limits are used as the alarm limits. Thus, for a data set that contains m data points of n process variables, a variable is said to violate the high alarm limit if it exceeds the high limit more than a critical number of times, LV_c. This critical number is calculated in a manner similar to the calculation of r_{95} in Eq. 9. In this case, $\epsilon = 0.025$ because high- and low-limit violations are considered to have the same probability. For $\epsilon = 0.025$ and $m = 1024$, $LV_c = 36$. Similarly, a variable is said to violate the low alarm limit if the variable is below the low limit more than LV_c times. For a historical data set, $\boldsymbol{H}$, which consists of m measurements of n variables, two n-dimensional vectors, $\alpha^{\boldsymbol{H}}$ and $\beta^{\boldsymbol{H}}$, are constructed:

$$\alpha^{\boldsymbol{H}}(i) = \begin{cases} 1 & \text{if the number of high-limit violations} \\ & \quad \text{for variable } i \text{ exceeds } LV_c \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

$$\beta^{\boldsymbol{H}}(i) = \begin{cases} 1 & \text{if the number of low-limit violations} \\ & \quad \text{for variable } i \text{ exceeds } LV_c \\ 0 & \text{otherwise,} \end{cases} \tag{14}$$

where $\alpha^{\boldsymbol{H}}(i)$ is the ith element of $\boldsymbol{\alpha}^{\boldsymbol{H}}$, and $\beta^{H}(i)$ is the ith element of $\boldsymbol{\beta}^{\boldsymbol{H}}$.

In order to compare the current snapshot data set $\boldsymbol{S}$ with a historical data set $\boldsymbol{H}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ vectors are constructed for each data set. Let $\boldsymbol{\alpha}^{\boldsymbol{S}}$ and $\boldsymbol{\beta}^{\boldsymbol{S}}$ denote the vectors for data set $\boldsymbol{S}$. Then the limit violation similarity factor is defined as

$$S_{LV} \triangleq \frac{1}{n} \sum_{i=1}^{n} N(i), \tag{15}$$

where

$$N(i) = \begin{cases} 1 & \text{if } \alpha^{S}(i) = \alpha^{\boldsymbol{H}}(i) \text{ and } \beta^{S}(i) = \beta^{\boldsymbol{H}}(i) \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

Thus, S_{LV} is the fraction of the n process variables that have similar limit violation behavior. Note that, $0 \leq S_{LV} \leq 1$. The two data sets $\boldsymbol{S}$ and $\boldsymbol{H}$ are considered to have similar patterns of limit violations if the value of S_{LV} exceeds a specified cutoff or threshold value.

The limit violation and PCA similarity factors can be utilized together to provide more effective pattern matching. When PCA models are developed, the process data are usually mean-centered and scaled to unit variance. Thus when data sets are compared using S_{PCA}, all information regarding the differences between the sample means is lost. But the S_{LV} metric captures this missing information for data sets where the mean shifts result in unusual alarm violations. The limit violation and PCA similarity factors can easily be combined because they both lie between zero and one.

Metrics for pattern matching

Two important metrics are proposed to quantify the effectiveness of a pattern-matching technique. First, however, several definitions are introduced:

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_1: The number of records in the candidate pool that are actually similar to the current snapshot, that is, the number of correctly identified records.

N_2: The total number of records in the candidate pool that are actually not similar to the current snapshot, that is, the number of incorrectly identified records. By definition, $N_1 + N_2 = N_P$.

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

The first metric, the *pool accuracy p*, characterizes the accuracy of the candidate pool

$$p \triangleq \frac{N_1}{N_P} \times 100\%. \tag{17}$$

A second metric, the *pattern matching efficiency* η, characterizes how effective the pattern-matching technique is in locating similar records in the historical database. It is defined as

$$\eta \triangleq \frac{N_1}{N_{DB}} \times 100\%. \tag{18}$$

When the pool size, N_P, is small ($N_P < N_{DB}$), then the efficiency η will be small because $N_1 \leq N_P$. A theoretical maximum efficiency, $\eta^{\max}$, for a given pool size N_P can be calculated as follows

$$\eta^{\max} \triangleq \begin{cases} \dfrac{N_P}{N_{DB}} \times 100\% & \text{for } N_P < N_{DB} \\ 100\% & \text{for } N_P \geq N_{DB}. \end{cases} \tag{19}$$

Because an effective pattern-matching technique should produce large values of both p and η, an average of the two quantities (ξ) is used as a measure of the overall effectiveness

$$\xi \triangleq \frac{p + \eta}{2}. \tag{20}$$

In pattern-matching problems, the relative importance of p and η metrics is application dependent. For example, suppose that a busy engineer wants to locate a small number (such as 2–5) of previous occurrences of an "abnormal plant operation" without having to waste time evaluating incorrectly identified records. In this situation, a large value of p is more important than a large value of η. In another application, it might be desirable to locate most (or even all) of previous occurrences of an abnormal plant operation, for business or legal reasons. Here, η is more important than p and a relatively large value of the candidate pool size, N_P, is acceptable. Fortunately, the proposed pattern-matching technique can accommodate both types of applications, as demonstrated in the following sections.

Remarks. The proposed pattern-matching methodology uses PCA to calculate the degree of similarity between multivariate time-series data sets. The biggest limitation of the proposed method is that a PCA model is a linear and static representation of the correlation between variables, while the data may represent nonlinear and dynamic processes. Thus, it is possible that for highly nonlinear and autocorrelated processes, the proposed methodology may not be very effective in matching patterns. However, in case studies of a highly nonlinear batch fermentation and the Tennessee Eastman Challenge Process, the standard PCA similarity factor provided very effective pattern matching (Singhal, 2002). Dynamic PCA (Ku et al., 1995) could be used to include the effects of autocorrelation, but our preliminary results have not indicated any significant improvement of dynamic PCA over the standard PCA similarity factor.

The methodology proposed in this article also assumes that the historical data are divided into periods of abnormal operation and that each disturbance begins at the beginning of the data window. In actual practice, it is likely that the historical data are neither preprocessed nor divided into data windows. Fortunately, the proposed pattern-matching approach is not very sensitive to the location of the start of the disturbance and the methodology can be extended to cases when the start and end times of disturbances are not known *a priori* (Singhal, 2002).

Case Study: Continuous Stirred-Tank Reactor

In order to compare different pattern-matching techniques, a case study was performed based on an extensive "historical database" for a simulated chemical reactor. A nonlinear continuous stirred-tank reactor (CSTR) with cooling-jacket dynamics, variable liquid level, and a first-order irreversible reaction, $A \rightarrow B$, was simulated. The CSTR and feedback control system are shown in Figure 2. Russo and Bequette (1996) derived a dynamic model for the CSTR based on the assumptions of perfect mixing and constant physical parameters. The mass, energy, and component balances

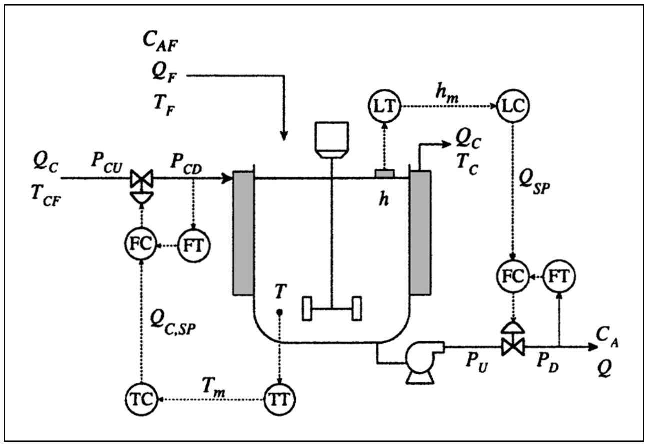

Figure 2. CSTR system with cascade control.

around the reactor and cooling jacket are

$$\frac{dC_A}{dt} = -k_0 e^{-E/RT} C_A + \frac{Q_F C_{AF} - Q C_A}{Ah} \tag{21}$$

$$\frac{dT}{dt} = \frac{k_0 e^{-E/RT} C_A(-\Delta H)}{\rho C_p} + \frac{(Q_F T_F - QT)}{Ah} + \frac{UA_C(T_C - T)}{\rho C_p Ah} \tag{22}$$

$$\frac{dT_C}{dt} = \frac{Q_C(T_{CF} - T_C)}{V_C} + \frac{UA_C(T - T_C)}{\rho_C C_{pC} V_C} \tag{23}$$

$$\frac{dh}{dt} = \frac{Q_F - Q}{A}, \tag{24}$$

where the process variables and model parameters are defined in the Notation section. The control valve dynamics are modeled by the first-order transfer function

$$G_v = \frac{K}{\tau s + 1}, \tag{25}$$

where $K = 1/16$ mA^{-1} and $\tau = 2$ s. The flow rate through each control valve is given by the following relation

$$\text{Flow rate} = C_v f(l) \sqrt{\frac{\Delta P_v}{g_s}}, \tag{26}$$

where the symbols are defined in the Notation section. A linear valve characteristic $(f(l) = l)$ is assumed, and the pressure drop across each control valve is assumed to be a constant for the entire flow range. The combination of these two assumptions results in a valve with a linear installed flow characteristic (Seborg et al., 1989). The control structure and the controller parameters have been described in detail by Johannesmeyer (1999). The nominal operating conditions and model parameters are given in Table 1.

In the simulation study, white noise is added to several measurements and process variables in order to simulate the variability present in real-world processes (Johannesmeyer, 1999).

Fault descriptions

Many fault-detection and diagnosis studies have been conducted using CSTR models (Vaidyanathan and Venkatasubramanian, 1992; Sorsa and Kovio, 1993), and a large number of possible fault conditions can be considered. The historical database for the present case study was designed to include both normal operating periods and a wide variety of abnormal situations or "faults." Setpoint changes in reactor temperature were also included. The 28 operating conditions are described in Table 2. Snapshot F_4 is identical to snapshot F4, except that the fault direction is negative, instead of positive.

The operating conditions in Table 2 include a wide range of disturbance and fault types that can be encountered in a typical historical database. The fault conditions include disturbances (e.g., ramp change in T_{CF}, step and sinusoidal disturbances in Q_F, etc.), instrumentation faults (e.g., dead coolant flow measurement, bias in reactor temperature measurement, etc.), and equipment faults (e.g., valve stiction, heat-exchanger fouling, catalyst deactivation, etc.). Ramp and step changes, oscillations and nonstationary disturbances, as well as valve stiction, are considered so that different fault characteristics could be included in the database.

Generation of historical database

In order to generate a large historical database, the CSTR system was simulated for a period of 39 days with measurements of the 14 process variables in Table 3 being recorded every 5 s. The last four measurements in Table 3 are the controller output signals. For example, hC is the signal in mA from the level controller and $Q_C C$ is the controller output in mA from the coolant flow-rate controller. Various pattern-matching techniques were evaluated for both the full- and reduced-measurement sets in Table 3.

The historical database was generated in the following manner. For each 120-min operating period, the mode of operation (that is, fault type, setpoint change, or normal operation) to be simulated was chosen randomly. The fault direction and magnitude were also randomly selected for each period. The fault direction could be positive or negative for faults that consisted of ramp or step changes. The fault magnitude was chosen randomly to be between 25% and 125% of the nominal value in Table 2. This means that *no two occurrences of any operating condition in the historical data are the same*. After the operation condition and any necessary parameters (i.e., direction and magnitude) were selected, the simulation ran for a total of 120 min. Each 120-min period was divided into 85.2 min for the selected operating mode and about 34.8 min for the process to return to the nominal steady state. Thus, faults occurred one at a time (no simultaneous faults) and had the same duration. Only the first 85.2 min of data for each operating period were retained in the historical database. The historical data consisted of 463 data sets (or data windows) where each data set contained 1024

Table 1. Nominal Operating Conditions and Model Parameters for the CSTR Case Study

$Q = 100$ L/min	$A = 0.1666$ m^2
$Q_C = 15$ L/min	$k_0 = 7.2\times10^{10}$ min^{-1}
$T_F = 320$ K	$\Delta H = -5\times10^4$ J/mol
$T_{CF} = 300$ K	$\rho C_p = 239$ J/(L·K)
$T = 402.35$ K	$\rho_C C_{pC} = 4175$ J/(L·K)
$T_C = 345.44$ K	$E/R = 8750$ K
$C_{AF} = 1.0$ mol/L	$UA_C = 5\times10^4$ J/(min·K)
$C_A = 0.037$ mol/L	$V_C = 10$ L
$h = 0.6$ m	

samples at 5-s intervals. This produced a total of over 474,000 data points for each of the 14 variables. Gaussian process and measurement noise was included in the simulation.

The historical database consists of 54 data windows for the normal operating conditions, and an average of 15 data windows for each of the other 27 operating conditions in Table 2. The amplitude of each fault, disturbance, or setpoint change was chosen randomly, as noted earlier. In this article, the disturbance, fault, or setpoint change always begins at the first sample in a historical data window, $\boldsymbol{H}_i$. However, current research has shown that the methodology is not very sensitive to the location of the start of the fault, and the proposed pattern-matching methodology can be extended to the more general situation where the start and end times of the operating conditions in the historical data are not known *a priori* (Singhal, 2002).

Generation of alarm limits

For the simulation study, 95% Shewhart chart limits were used as the alarm limits for the S_{LV} calculations. The chart limits were constructed using "representative" data that included small disturbances. The representative operating peri-

Table 2. Operating Conditions and Faults

ID	Operating Condition	Description	Nominal Value
N	Normal operation	Operation at the nominal conditions. No disturbances	N/A
F1	Catalyst deactivation	The activation energy ramps up	The ramp rate for E/R is $+3$ K/min
F2	Heat-exchanger fouling	The heat-transfer coefficient ramps down	The ramp rate for U_{AC} is -125 [J/min(K)]/min
F3	Dead-coolant flow measurement	The coolant flow-rate measurement stays at its last value	N/A
F4, F_4	Bias in reactor temperature measurement	The reactor temperature measurement has a bias	± 4 K
F5, F_5	Coolant valve stiction +F7	Dead band for stiction = 5% of the valve span	N/A
F6, F_6	Step change in Q_F	Step change in feed flow rate	± 10 L/min
F7, F_7	Ramp change in C_{AF}	The feed concentration ramps up or down	The ramp rate is $\pm 6\times10^{-4}$(mol/L)/min
F8, F_8	Ramp change in T_F	The feed temperature ramps up or down	The ramp rate is ± 0.1 K/min
F9, F_9	Ramp change in T_{CF}	The coolant feed temperature ramps up or down	The ramp rate is ± 0.1 K/min
F10, F_10	Step change in P_{CU}	Step change in upstream pressure in the cooling line	± 2.5 psi
F11, F_11	Step change in P_D	Step change in downstream pressure in the reactor outlet line	± 5 psi
F12	Damped oscillations in feed flow rate	The feed flow changes as: $e^{-t/33}\sin(2\pi t/10)$ L/min	10 L/min
F13	Autoregressive disturbance in feed flow rate	$Q_F(k) = 0.8\times Q_F(k-1) + w(k)$. $w(k) \sim N(0,1)$	N/A
S1, S_1	Setpoint change in T	Setpoint change for the reactor temperature	± 3 K
O1	High-frequency oscillations in feed flow rate	Sinusoidal oscillations of frequency 3 cycles/min	10 L/min
O2	Intermediate frequency oscillations in feed flow rate	Sinusoidal oscillations of frequency 1 cycle/min	10 L/min
O3	Intermediate frequency oscillations in feed flow rate	Sinusoidal oscillations of frequency 0.5 cycle/min	10 L/min
O4	Low-frequency oscillations in feed flow rate	Sinusoidal oscillations of frequency 0.2 cycle/min	10 L/min

Table 3. Measurement Sets for the CSTR Case Study

Variable	Reduced Set	Full Set
C_A		√
T	√	√
T_C		√
h	√	√
Q	√	√
Q_C	√	√
Q_F	√	√
C_{AF}		√
T_F		√
T_{CF}		√
hC	√	√
QC	√	√
TC	√	√
$Q_C C$	√	√

ods are described in Table 4. Except for run N1, each run includes two disturbances, one in the positive direction and one in the negative direction. Thus, a total of 13 representative operating periods were considered, with each period lasting 120 min. The high and low limits for each variable were calculated using these data (Johannesmeyer, 1999). Although Shewhart chart limits were used for this case study, the S_{LV} alarm limits could also be specified by the plant operators.

Data analysis

The pattern-matching analysis began by choosing one of the operating conditions in Table 2 as the current snapshot, $\boldsymbol{S}$. For all pattern-matching techniques, except the limit violation similarity factor, the snapshot data $\boldsymbol{S}$ were scaled to zero mean and unit variance. Then, each historical data record $\boldsymbol{H}_i$ was scaled using the scaling factors for the snapshot data. This scaling is consistent with the problem statement of "finding patterns similar to the current snapshot of data." After the historical data $\boldsymbol{H}_i$ were scaled, the pattern-matching calculations were performed and metrics p, η, and ξ were calculated.

Table 4. Representative Operating Periods Used to Determine Shewhart Chart Limits

Run	Run Description
N1	No unusual disturbances
N2, N_2	Sinusoidal pulse (half-period) in coolant feed temperature with an amplitude of ±3 K
N3, N_3	Sinusoidal pulse in feed temperature with an amplitude of ±3 K
N4, N_4	Sinusoidal pulse in downstream pressure in the reactor outlet line with an amplitude of ±1.5 psi
N5, N_5	Rectangular pulse in upstream pressure in the coolant line with an amplitude of ±0.75 psi (a step up at 250 min and a step back down at 1000 min)
N6, N_6	Feed flow-rate ramps at a rate of ±0.0075 L/min over 400 min and stays there for 800 min
N7, N_7	First-order exponential change ($\tau = 100$ min) in feed concentration to a new steady-state value of 1 ± 0.018 mol/L

This analysis was repeated for 28 different snapshots, one for each one of the 28 operating conditions in Table 2. The snapshots were based on the nominal conditions shown in Table 2, but the historical data sets had random fault magnitudes, as described earlier.

Averaging of data

Large quantities of historical data are averaged in order to "compress" the data and to reduce the degree of autocorrelation. In order to evaluate the effects of averaging historical data, 2-min averages of the original 5-s sampled data were also considered. Thus, each window of averaged data contained 42 observations for each measured variable. The effect of data averaging on pattern matching was investigated by comparing results for the original 5-s data and the 2-min averaged data.

Results and Discussion

A variety of pattern-matching methods have been compared based on the pool accuracy and pattern-matching efficiency metrics of Eqs. 17 and 18. The average values of p, η, and ξ for the 28 operating conditions in Table 2 were used as the basis for these comparisons. The size of the candidate pool, N_P, was also an important consideration. Simulations were performed for both the full- and the reduced-measurement sets for the CSTR case study, and for both the original 5-s data and 2-min averaged data. The proposed pattern-matching methodology has also been successfully evaluated in case studies involving a highly nonlinear acetone−butanol batch fermentation system and the Tennessee Eastman Challenge Process (Singhal, 2002).

Comparison of pattern-matching techniques

The following pattern-matching techniques were evaluated in the CSTR case study: T^2, Q, combined discriminant, PCA, and limit violation similarity methods. In order to determine the best performance of the PCA and limit violation similarity factors, the S_{PCA} cutoff value was gradually increased from 0 to 1, and the average values of p, η, and ξ were calculated for the 28 snapshot data sets. Representative results are presented in Figure 3. As the S_{PCA} cutoff value increases, η decreases and N_P decreases, as will be demonstrated later. The value of p first increases with an increase in the cutoff value, but then decreases as the cutoff value approaches unity. This result occurs because the value of p that is plotted is the average value for the 28 operating conditions. Thus, when the S_{PCA} cutoff becomes sufficiently high, the candidate pool for some operating conditions becomes very small, and in some cases, zero. This reduction leads to a lower average value of p. The average of p and η, ξ, is less sensitive to the cutoff value and exhibits a broad maximum. The cutoff value that maximizes ξ is considered to be the optimum cutoff.

When S_{PCA} and S_{LV} are used together, data sets $\boldsymbol{S}$ and $\boldsymbol{H}_i$ are considered to be similar if both similarity factors exceed their cutoff values. The optimum cutoffs for this S_{PCA}-S_{LV} method were obtained by plotting the value of ξ vs. the S_{PCA} and S_{LV} cutoffs as a 3-D surface. The location of the highest

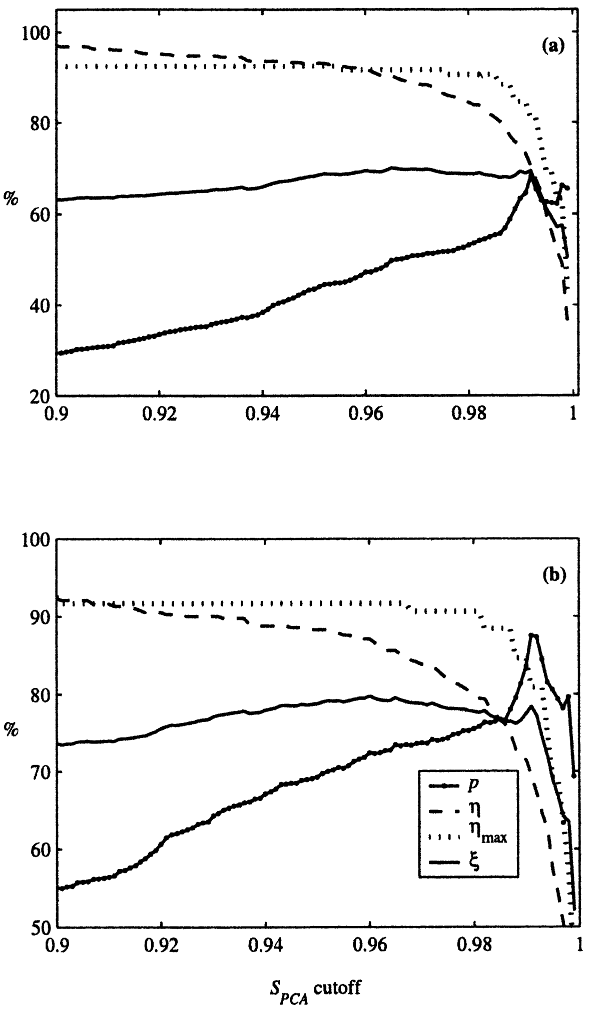

Figure 3. Effect of PCA similarity factor cutoff on pattern matching.

Full measurement set and 5-s data. Results for: (a) S_{PCA}, (b) S_{PCA}-S_{LV} (S_{LV} cutoff = 0.660).

point on the surface was taken as the location optimum cutoffs for the two similarity factors. The optimum cutoffs for S_{PCA} and S_{LV} using this method were found to be 0.960 and 0.660, respectively. Figure 3b shows the variations of the average value of p, η, and ξ, with the S_{PCA} cutoff when the S_{LV} cutoff is assigned its optimum value of 0.660.

Table 5 summarizes the best performance of each technique for the full measurement set and 5-s data. The T^2, Q, combined discriminant and S_{LV} methods produce very low p values due to their large candidate pool sizes. Also, the corresponding ξ values are quite low compared to those obtained for the S_{PCA} method. The S_{PCA} method produces relatively small values of N_P. The S_{PCA} method correctly identifies half of the data sets in the candidate pool while retrieving over 90% of the similar data sets in the historical database. Although the S_{LV} method has a relatively poor performance by itself, its combination with S_{PCA} provides significantly improved results. The p value for the S_{PCA}-S_{LV} method is almost 45% better than for the S_{PCA} method alone while producing only a marginal decrease in η. The results in Table 5 indicate that S_{PCA}-S_{LV} method provides the best performance, based on the value of ξ.

Table 5. Best Performance for the Full Measurement Set and 5-s Data

Method	Best Cutoff(s)	N_P	p (%)	η (%)	$\eta_{\max}$ (%)	ξ (%)
T^2 statistic (95% limit)	N/A	103	27	69	100	**48**
Q statistic (95% limit)	N/A	56	19	52	100	**35**
Combined Q and T^2	N/A	86	18	86	100	**51**
PCA similarity factor, S_{PCA}	0.965	34	50	90	99	**70**
Limit violation similarity factor, S_{LV}	0.660	83	25	94	100	**60**
PCA and LV similarity factors	**0.960, 0.660**	**20**	**72**	**87**	**95**	**80**

The detailed results for the S_{PCA}-S_{LV} method and the 28 different operating conditions are presented in Table 6. The misclassification details for the 28 snapshots are shown in Table 7. The number of misclassifications for each snapshot is N_2, where $N_2 = N_P - N_1$. Some interesting patterns can be seen. For example, operating condition F4 (bias in reactor temperature measurement) is misclassified with operating conditions F2 (heat exchanger fouling) and S_1 (negative setpoint change in T). All three of these operating conditions result in an increase in the coolant flow rate, Q_C, and a decrease in the coolant temperature, T_C. Consequently, they

Table 6. Results for the S_{PCA}-S_{LV} Method for 5-s Data and Full Measurement Set

Snapshot	N_{DB}	N_1	N_2	p (%)	η (%)	$\eta_{\max}$ (%)	ξ (%)
Average	**17**	**15**	**6**	**72**	**87**	**100**	**80**
Normal	54	54	14	79	100	100	90
F1	23	23	0	100	100	100	100
F2	32	32	11	74	100	100	87
F3	30	30	0	100	100	100	100
F4	14	8	19	30	57	100	43
F_4	8	6	8	43	75	100	59
F5	15	14	3	82	93	100	88
F_5	16	16	0	100	100	100	100
F6	15	15	0	100	100	100	100
F_6	14	13	0	100	93	93	96
F7	15	15	0	100	100	100	100
F_7	12	12	0	100	100	100	100
F8	15	15	3	83	100	100	92
F_8	20	20	0	100	100	100	100
F9	9	9	8	53	100	100	76
F_9	12	12	8	60	100	100	80
F10	15	15	14	52	100	100	76
F_10	14	14	15	48	100	100	74
F11	14	10	5	67	71	100	69
F_11	14	6	11	35	43	100	39
S1	9	9	6	60	100	100	80
S_1	14	14	14	50	100	100	75
F12	12	1	3	25	8	33	17
F13	12	5	4	56	42	75	49
O1	19	17	0	100	89	89	95
O2	12	11	2	85	92	100	88
O3	12	12	2	86	100	100	93
O4	12	9	6	60	75	100	68

Note: PCA cutoff = 0.960; LV cutoff = 0.660.

Table 7. Misclassification Results for Table 6

Snapshot	Pool Size (N_P)	Correct (N_1)	Misclassifications (fault and No.)
Normal	68	54	F4 (3), F_4 (2), F_10 (1), F11 (1), F_11 (1), S_1 (1), F13 (5)
F1	23	23	None
F2	43	32	F4 (9), S_1 (2)
F3	30	30	None
F4	27	8	F2 (8), S_1 (11)
F_4	14	6	S1 (8)
F5	17	14	F8 (3)
F_5	16	15	None
F6	15	15	None
F_6	13	13	None
F7	15	15	None
F_7	12	12	None
F8	18	15	F5 (3)
F_8	20	20	None
F9	17	9	F_9 (8)
F10	29	15	F_10 (14)
F_10	29	14	F10 (15)
F11	15	10	F_11 (5)
F_11	17	6	F11 (11)
F12	4	1	O4 (3)
F13	9	5	O2 (1), O3 (1), O4 (2)
S1	15	9	F_4 (6)
S_1	28	14	F4 (14)
O1	17	17	None
O2	13	11	F13 (2)
O3	14	12	F13 (1), O2 (1)
O4	15	9	F12 (6)

tend to appear in the same candidate pool, and it is difficult for the pattern-matching techniques to distinguish between them. Also, it is difficult for pattern-matching techniques to distinguish between positive and negative directions of faults F10 and F11. However, the distinction between the positive and negative fault directions may or may not be important, depending on the application.

In order to develop further insight into misclassification issues associated with the CSTR case study, similarity factors were calculated for possible pairs of the nominal operating conditions in Table 2. These results are shown in Tables 8 and 9, respectively. For each row, the operating condition in the first column is considered to be the snapshot for scaling. The off-diagonal values of the similarity factors in Tables 8 and 9 that are greater than their respective cutoff values of 0.960 (for S_{PCA}) and 0.660 (for S_{LV}) are in **bold** face type. Note that Table 8 exhibits a diagonal dominance, while Table 9 shows less structure.

If only the S_{PCA} is used, then Table 8 shows that the positive and negative changes for operating conditions F4 through F11 have a high possibility of being misclassified with each other. But pairs of operating conditions like F4 and F_4 have small S_{LV} values. Consequently, using S_{LV} and S_{PCA} together helps to distinguish between these positive and negative changes. In fact, the combination of similarity factors provides a significant improvement over using S_{PCA} alone, as shown in Table 5.

The results presented in Tables 8 and 9 for the nominal values of the operating condition provide considerable insight into the misclassification problem. For example, operating conditions F2, F4 and S_1 show similarities because these results in an increase in the coolant flow rate, Q_C, and a decrease in the coolant temperature, T_C, as mentioned earlier. Operating conditions F6 and F12 show similarities using S_{PCA} only, but are not similar when both S_{PCA} and S_{LV} are used together. Thus, these tables indicate the operating conditions that one might expect to be misclassified because of their similar patterns. But, of course, the random fault and disturbance magnitudes used in the simulations can result in other operating conditions being misclassified with each other, as shown in Table 7.

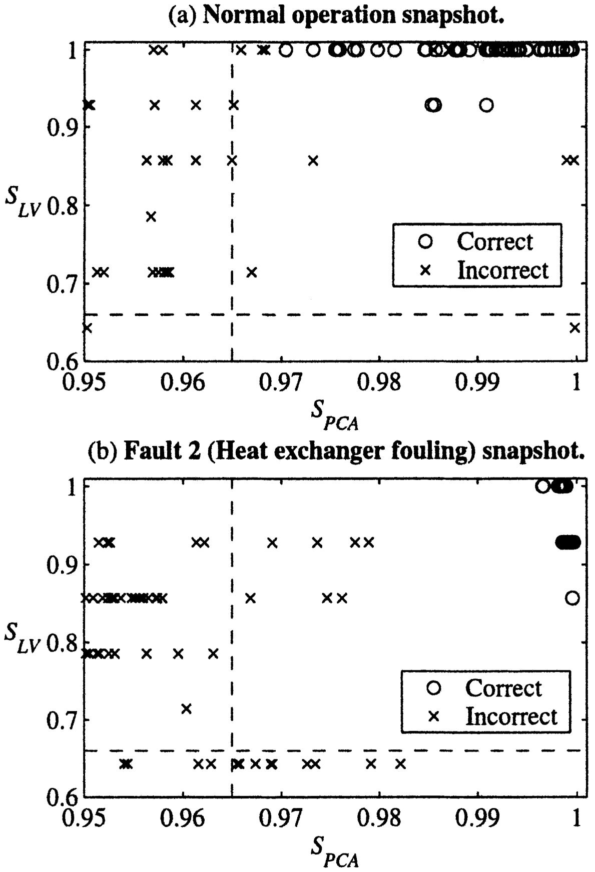

Figure 4. Scatter plots of the PCA and LV similarity factors for 5-s data and the full measurement set.

The best cutoffs are shown as dashed lines.

For all of the previous results, the candidate pool was selected by specifying cutoff values for the PCA and limit violation similarity factors. But in actual applications, the best cutoff values are not known *a priori*. Thus, other ways of determining the candidate pool must be used. Two such strategies are now proposed.

Selection of a candidate pool using scatter plots

It is useful to plot the calculated S_{PCA} and S_{LV} values for a particular snapshot data set and all of the historical data sets on a scatter diagram. Representative scatter diagrams for two snapshots are shown in Figure 4 for the historical data sets that have the largest values of S_{PCA} or S_{LV}. Each

Table 8. PCA Similarity Factor for the Nominal Values of the Operating Conditions

Op ID	N	F1	F2	F3	F4	F_4	F5	F_5	F6	F_6	F7	F_7	F8	F_8	F9	F_9	F10	F_10	F11	F_11	S1	S_1	F12	F13	O1	O2	O3	O4
N	*1*	0.93	0.83	0.31	0.83	0.82	0.59	0.59	0.88	0.85	0.95	0.92	**0.96**	**0.96**	0.82	0.80	0.80	0.79	0.90	0.90	0.83	0.86	0.86	0.83	0.79	0.76	0.80	0.86
F1	0.93	*1*	0.96	0.34	0.94	0.93	0.77	0.71	0.86	0.81	0.83	0.83	0.85	0.85	0.84	0.83	0.85	0.84	0.80	0.80	0.92	0.92	0.82	0.79	0.80	0.80	0.80	0.81
F2	0.80	0.83	*1*	0.37	0.87	0.88	0.81	0.81	0.74	0.73	0.81	0.81	0.81	0.81	0.81	0.81	0.86	0.85	0.80	0.80	0.82	0.82	0.74	0.79	0.80	0.74	0.78	0.79
F3	0.55	0.65	0.66	*1*	0.67	0.67	0.65	0.65	0.51	0.51	0.52	0.52	0.52	0.52	0.52	0.52	0.59	0.59	0.52	0.52	0.67	0.67	0.51	0.51	0.52	0.50	0.50	0.51
F4	**0.97**	0.81	**0.96**	0.34	*1*	**> 0.99**	0.82	0.82	0.87	0.82	0.81	0.81	0.82	0.82	0.82	0.81	0.85	0.84	0.80	0.80	**0.98**	**0.98**	0.84	0.79	0.79	0.78	0.79	0.82
F_4	**0.96**	0.81	**0.96**	0.34	**> 0.99**	*1*	0.82	0.82	0.88	0.83	0.81	0.81	0.83	0.83	0.82	0.82	0.85	0.84	0.80	0.80	**0.97**	**0.97**	0.84	0.79	0.80	0.78	0.78	0.82
F5	0.79	0.79	0.81	0.47	0.80	0.80	*1*	**> 0.99**	0.73	0.70	0.70	0.70	0.84	0.84	0.72	0.71	0.83	0.84	0.70	0.72	0.80	0.80	0.74	0.71	0.68	0.68	0.68	0.77
F_5	0.78	0.80	0.80	0.47	0.80	0.80	**> 0.99**	*1*	0.73	0.70	0.70	0.70	0.83	0.83	0.71	0.71	0.81	0.82	0.70	0.72	0.79	0.79	0.74	0.71	0.67	0.67	0.68	0.77
F6	0.75	0.74	0.84	0.35	0.85	0.85	0.73	0.73	*1*	**> 0.99**	0.72	0.73	0.73	0.73	0.73	0.72	0.76	0.75	0.75	0.75	0.81	0.81	**> 0.99**	0.76	0.75	0.74	0.75	0.90
F_6	0.74	0.74	0.84	0.35	0.86	0.86	0.73	0.73	**> 0.99**	*1*	0.72	0.73	0.73	0.73	0.73	0.72	0.76	0.75	0.75	0.75	0.82	0.82	**> 0.99**	0.77	0.75	0.74	0.75	0.90
F7	0.87	0.84	0.95	0.28	0.92	0.92	0.83	0.83	0.80	0.77	*1*	**> 0.99**	0.83	0.83	0.83	0.82	0.85	0.84	0.80	0.80	0.88	0.88	0.78	0.79	0.80	0.74	0.78	0.81
F_7	0.90	0.83	0.95	0.28	0.93	0.93	0.83	0.83	0.81	0.78	**> 0.99**	*1*	0.84	0.83	0.83	0.82	0.85	0.84	0.80	0.80	0.89	0.89	0.79	0.79	0.80	0.74	0.78	0.81
F8	0.91	0.86	0.95	0.27	0.92	0.92	0.84	0.83	0.80	0.77	0.82	0.82	*1*	**> 0.99**	0.83	0.83	0.85	0.84	0.80	0.80	0.88	0.89	0.79	0.79	0.80	0.74	0.78	0.81
F_8	0.92	0.85	0.96	0.27	0.93	0.93	0.83	0.83	0.81	0.77	0.82	0.82	**> 0.99**	*1*	0.83	0.83	0.85	0.85	0.80	0.80	0.89	0.89	0.79	0.79	0.80	0.74	0.78	0.81
F9	0.92	0.84	0.91	0.46	0.86	0.86	0.84	0.84	0.83	0.83	0.83	0.83	0.84	0.84	*1*	**> 0.99**	0.88	0.88	0.83	0.83	0.93	0.92	0.83	0.82	0.83	0.81	0.82	0.83
F_9	0.93	0.85	0.93	0.46	0.87	0.87	0.85	0.85	0.83	0.83	0.84	0.84	0.84	0.84	**> 0.99**	*1*	0.88	0.88	0.83	0.83	0.94	0.93	0.84	0.82	0.83	0.81	0.82	0.83
F10	0.85	0.83	0.84	0.52	0.84	0.84	0.74	0.75	0.82	0.80	0.83	0.83	0.82	0.82	0.84	0.84	*1*	**> 0.99**	0.83	0.83	0.84	0.84	0.83	0.80	0.83	0.78	0.79	0.77
F_10	0.84	0.83	0.84	0.53	0.84	0.84	0.74	0.76	0.82	0.81	0.83	0.83	0.82	0.82	0.84	0.84	**> 0.99**	*1*	0.83	0.83	0.84	0.84	0.83	0.80	0.83	0.79	0.79	0.80
F11	0.84	0.83	0.83	0.51	0.83	0.83	0.79	0.79	0.83	0.83	0.82	0.83	0.82	0.82	0.80	0.80	0.83	0.82	*1*	**0.99**	0.84	0.84	0.82	0.82	0.83	0.80	0.81	0.82
F_11	0.85	0.83	0.84	0.51	0.85	0.85	0.79	0.79	0.83	0.83	0.82	0.83	0.83	0.83	0.81	0.81	0.82	0.82	**0.99**	*1*	0.85	0.85	0.82	0.82	0.83	0.80	0.81	0.82
S1	0.92	0.82	0.94	0.32	**0.98**	**0.98**	0.81	0.82	0.83	0.82	0.82	0.82	0.83	0.83	0.83	0.82	0.84	0.84	0.80	0.80	*1*	**> 0.99**	0.85	0.79	0.80	0.78	0.79	0.81
S_1	0.76	0.75	0.93	0.25	**0.96**	0.96	0.81	0.82	0.80	0.70	0.77	0.77	0.78	0.78	0.78	0.77	0.84	0.84	0.65	0.65	**> 0.99**	*1*	0.81	0.79	0.71	0.73	0.73	0.76
F12	0.83	0.76	0.85	0.39	0.92	0.92	0.66	0.65	**0.99**	0.99	0.72	0.72	0.74	0.74	0.71	0.71	0.73	0.72	0.73	0.75	0.91	0.90	*1*	0.84	0.74	0.79	0.84	**0.98**
F13	0.82	0.81	0.80	0.43	0.83	0.82	0.64	0.64	0.84	0.83	0.73	0.75	0.74	0.74	0.76	0.78	0.76	0.77	0.76	0.74	0.82	0.82	0.84	*1*	0.90	0.96	**0.98**	0.83
O1	0.83	0.82	0.83	0.53	0.83	0.83	0.74	0.74	0.83	0.83	0.80	0.81	0.81	0.82	0.81	0.81	0.81	0.81	0.83	0.82	0.83	0.83	0.83	0.83	*1*	0.82	0.83	0.83
O2	0.71	0.70	0.69	0.24	0.70	0.69	0.41	0.40	0.76	0.75	0.68	0.69	0.69	0.68	0.72	0.71	0.72	0.72	0.69	0.76	0.69	0.69	0.76	0.91	0.84	*1*	0.91	0.77
O3	0.48	0.50	0.47	0.22	0.47	0.46	0.40	0.39	0.78	0.75	0.36	0.36	0.38	0.38	0.38	0.37	0.67	0.42	0.39	0.62	0.46	0.47	0.85	0.87	0.67	0.92	*1*	0.84
O4	0.47	0.53	0.62	0.34	0.62	0.61	0.37	0.37	0.80	0.80	0.36	0.36	0.38	0.38	0.37	0.36	0.66	0.42	0.39	0.40	0.61	0.61	**0.98**	0.94	0.34	0.66	0.79	*1*

Note: Similarity factor values greater than the cutoff of 0.960 are in bold face type.

Table 9. Limit Violation Similarity Factor for the Nominal Values of the Operating Conditions

Op ID	N	F1	F2	F3	F4	F_4	F5	F_5	F6	F_6	F7	F_7	F8	F_8	F9	F_9	F10	F_10	F11	F_11	S1	S_1	F12	F13	O1	O2	O3	O4
N	*1*	0.64	0.64	0.57	0.57	0.57	0.64	0.64	0.21	0.21	0.57	0.64	0.64	0.64	**0.71**	**0.71**	**0.93**	**0.93**	**0.93**	**0.93**	0.57	0.57	0.21	0.57	0.71	0.50	0.29	0.21
F1	0.64	*1*	0.57	0.57	0.64	**0.86**	0.57	**0.86**	0.29	0.50	0.64	**0.86**	0.57	**0.86**	0.57	**0.79**	0.71	0.64	0.57	0.57	**0.86**	0.64	0.29	0.36	0.43	0.21	0.29	0.21
F2	0.64	0.57	*1*	0.64	**0.93**	0.57	**0.86**	0.57	0.57	0.21	**0.79**	0.57	**0.86**	0.57	**0.79**	0.57	0.64	**0.71**	0.57	0.57	0.64	**0.86**	0.50	0.29	0.43	0.21	0.21	0.21
F3	0.57	0.57	0.64	*1*	0.64	0.57	0.57	0.50	0.29	0.21	0.57	0.50	0.57	0.50	0.57	0.50	0.57	0.57	0.50	0.50	0.57	0.64	0.36	0.36	0.57	0.36	0.43	0.57
F4	0.57	0.64	**0.93**	0.64	*1*	0.57	**0.79**	0.50	0.64	0.21	**0.86**	0.50	**0.79**	0.50	**0.71**	0.50	0.57	0.64	0.50	0.50	0.64	**0.93**	0.57	0.36	0.43	0.21	0.21	0.21
F_4	0.57	**0.86**	0.57	0.57	0.57	*1*	0.50	0.21	0.64	0.50	**0.79**	0.50	**0.79**	0.50	**0.71**	0.64	0.57	0.50	0.50	**0.50**	**0.93**	0.64	0.21	0.29	0.43	0.21	0.29	0.21
F5	0.64	0.57	**0.86**	0.57	**0.79**	0.50	*1*	0.64	0.43	0.14	**0.79**	0.57	**1.00**	0.64	**0.79**	0.57	0.64	0.71	0.57	0.57	0.50	**0.79**	0.43	0.21	0.36	0.14	0.14	0.14
F_5	0.64	**0.86**	0.57	0.50	0.50	**0.79**	0.64	*1*	0.14	0.43	0.50	**0.86**	0.64	1.00	0.57	**0.79**	**0.71**	0.64	0.57	0.57	**0.79**	0.50	0.14	0.21	0.36	0.14	0.21	0.14
F6	0.21	0.29	0.57	0.29	0.64	0.21	0.43	0.14	*1*	0.21	0.50	0.14	0.43	0.14	0.36	0.14	0.21	0.29	0.29	0.21	0.29	0.57	0.57	0.29	0.21	0.21	0.21	0.21
F_6	0.21	0.50	0.21	0.21	0.21	0.64	0.14	0.43	0.21	*1*	0.14	0.43	0.14	0.43	0.14	0.36	0.29	0.21	0.21	0.29	0.57	0.29	0.21	0.21	0.21	0.21	0.29	0.21
F7	0.57	0.64	**0.79**	0.57	0.86	0.50	**0.79**	0.50	0.50	0.14	*1*	0.57	**0.79**	0.50	0.71	0.50	0.57	0.64	0.50	0.50	0.50	**0.86**	0.50	0.29	0.36	0.14	0.14	0.14
F_7	0.64	**0.86**	0.57	0.50	0.50	**0.79**	0.57	**0.86**	0.14	0.43	0.57	*1*	0.57	**0.86**	0.57	**0.79**	**0.71**	0.64	0.57	0.57	**0.79**	0.50	0.14	0.21	0.36	0.14	0.21	0.14
F8	0.64	0.57	**0.86**	0.57	0.79	0.50	**1.00**	0.64	0.43	0.14	**0.79**	0.57	*1*	0.64	**0.79**	0.57	0.64	**0.71**	0.57	0.57	0.50	**0.79**	0.43	0.21	0.36	0.14	0.14	0.14
F_8	0.64	0.86	0.57	0.50	0.50	0.79	0.64	1.00	0.14	0.43	0.50	0.86	0.64	*1*	0.57	0.79	0.71	0.64	0.57	0.57	0.79	0.50	0.14	0.21	0.36	0.14	0.21	0.14
F9	**0.71**	0.57	**0.79**	0.57	**0.71**	0.50	**0.79**	0.57	0.36	0.14	**0.71**	0.57	**0.79**	0.57	*1*	**0.71**	**0.71**	**0.79**	0.64	0.64	0.50	**0.71**	0.36	0.29	0.43	0.21	0.21	0.14
F_9	**0.71**	**0.79**	0.57	0.50	0.50	**0.71**	0.57	**0.79**	0.14	0.36	0.50	**0.79**	0.57	**0.79**	**0.71**	*1*	**0.79**	**0.71**	0.64	0.64	**0.71**	0.50	0.14	0.29	0.43	0.21	0.29	0.14
F10	**0.93**	**0.71**	0.64	0.57	0.57	0.64	0.64	**0.71**	0.21	0.29	0.57	**0.71**	0.64	**0.71**	**0.71**	**0.79**	*1*	**0.93**	**0.86**	**0.86**	0.64	0.57	0.21	0.50	0.64	0.43	0.36	0.21
F_10	**0.93**	0.64	**0.71**	0.57	0.64	0.57	**0.71**	0.64	0.29	0.21	0.64	0.64	**0.71**	0.64	**0.79**	**0.71**	**0.93**	*1*	**0.86**	**0.86**	0.57	0.64	0.29	0.50	0.64	0.43	0.29	0.21
F11	**0.93**	0.57	0.57	0.50	0.50	0.50	0.57	0.57	0.29	0.21	0.50	0.57	0.57	0.57	0.64	0.64	**0.86**	**0.86**	*1*	**0.93**	0.50	0.50	0.21	0.50	0.64	0.50	0.29	0.21
F_11	**0.93**	0.57	0.57	0.50	0.50	0.50	0.57	0.57	0.21	0.29	0.50	0.57	0.57	0.57	0.64	0.64	**0.86**	**0.86**	**0.93**	*1*	0.50	0.50	0.21	0.50	0.64	0.50	0.29	0.21
S1	0.57	**0.86**	0.64	0.57	0.64	**0.93**	0.50	**0.79**	0.29	0.57	0.50	**0.79**	0.50	**0.79**	0.50	**0.71**	0.64	0.57	0.50	0.50	*1*	0.57	0.21	0.29	0.43	0.21	0.29	0.21
S_1	0.57	0.64	**0.86**	0.64	**0.93**	0.64	**0.79**	0.50	0.57	0.29	**0.86**	0.50	**0.79**	0.50	**0.71**	0.50	0.57	0.64	0.50	0.50	0.57	*1*	0.57	0.36	0.43	0.21	0.21	0.21
F12	0.21	0.29	0.50	0.36	0.57	0.21	0.43	0.14	0.57	0.21	0.50	0.14	0.43	0.14	0.36	0.14	0.21	0.29	0.21	0.21	0.21	0.57	*1*	0.64	0.43	0.64	0.64	0.64
F13	0.57	0.36	0.29	0.36	0.36	0.29	0.21	0.21	0.29	0.21	0.29	0.21	0.21	0.21	0.29	0.29	0.50	0.50	0.50	0.50	0.29	0.36	0.64	*1*	0.79	0.86	0.64	0.57
O1	**0.71**	0.43	0.43	0.57	0.43	0.43	0.36	0.36	0.21	0.21	0.36	0.36	0.36	0.36	0.43	0.43	0.64	0.64	0.64	0.64	0.43	0.43	0.43	**0.79**	*1*	**0.79**	0.57	0.50
O2	0.50	0.21	0.21	0.36	0.21	0.21	0.14	0.14	0.21	0.21	0.14	0.14	0.14	0.14	0.21	0.21	0.43	0.43	0.50	0.50	0.21	0.21	0.64	**0.86**	**0.79**	*1*	**0.79**	**0.71**
O3	0.29	0.29	0.21	0.43	0.21	0.29	0.14	0.21	0.21	0.29	0.14	0.21	0.14	0.21	0.21	0.29	0.36	0.29	0.29	0.29	0.29	0.21	0.64	0.64	0.57	**0.79**	*1*	**0.86**
O4	0.21	0.21	0.21	0.57	0.21	0.21	0.14	0.14	0.21	0.21	0.14	0.14	0.14	0.14	0.14	0.14	0.21	0.21	0.21	0.21	0.21	0.21	0.64	0.57	0.50	**0.71**	**0.86**	*1*

Note: Similarity factor values greater than the cutoff of 0.660 are in bold face type.

Table 10. Results for a Specified Pool Size Using the S_{PCA}-S_{LV} Method for the Full-Measurement Set and 5-s Data

	Candidate Pool Size (N_P)									
	5		10		15		20		25	
Snapshot	p (%)	η (%)	p (%)	η (%)	p (%)	η (%)	p (%)	η (%)	p (%)	η (%)
N	100	9	100	19	100	28	100	37	100	46
F2	100	16	100	31	100	47	100	63	100	78
F4	100	36	70	50	47	50	55	79	44	79
F10	100	33	90	60	60	60	50	67	52	87
F_11	100	36	90	64	73	79	65	93	52	93
F13	60	25	40	33	27	33	20	33	16	33
O4	80	33	40	33	53	67	50	83	48	100
Average (all)	**92**	**32**	**86**	**59**	**76**	**76**	**63**	**82**	**54**	**86**

data point in the scatter diagram represents a data window, $\boldsymbol{H}_i$, in the historical database. For clarity, the data windows that actually are previous occurrences of the snapshot data are marked as ○, and other operating periods are marked as ×. In practice, the user will not be able to distinguish between the ○ and × data sets because these distinctions are not known. Thus, a reasonable approach would be to select data windows that are close to the (1,1) point to be the records in the "candidate pool." This approach allows the user to manually cluster the data on the scatter diagram and select N_P and the candidate pool accordingly.

Selection of a candidate pool from a rank-ordered list

In this approach, the similarity factor(s) for each of the $\boldsymbol{H}_i$ data windows and snapshot $\boldsymbol{S}$ are sorted in decreasing order. If more than one similarity factor is used, a weighted average of the similarity factors can be calculated as

$$SF = \phi S_{PCA} + (1-\phi) S_{LV}, \tag{27}$$

where, $\phi(0 \le \phi \le 1)$, is the weighting factor. In our experience, a value of $\phi = 0.5$ is satisfactory. Thus, equal weight is given to each similarity factor.

After the rank-ordered list of promising historical data windows has been constructed, the individual data windows are examined starting with the first record, because it has the largest SF value. This process continues until the user is satisfied with the results. If the user wishes to locate only a few historical data windows that are very similar to the current snapshot, then a relatively small number of records need be examined. However, if the user is interested in locating almost all of the previous occurrences of the current abnormal operation, there is motivation to examine a much larger number of records. For example, this type of situation could occur after a plant accident or a product recall.

In this approach, neither the cutoff valve nor the size of the candidate pool, N_P, have to be specified. But if N_P is specified, the candidate pool can be generated automatically. This option eliminates the subjective judgment that must be exercised if the candidate pool is formed by inspecting a scatter diagram.

An example of this approach is shown in Table 10, where the pool accuracy and pattern-matching efficiency for different candidate pool sizes are analyzed for a few representative operating conditions. For very small values of N_P, the pool accuracy is large because only the most similar $\boldsymbol{H}_i$ are selected for the candidate pool. As N_P is increased, p decreases and η increases because more records are being included in the candidate pool. This analysis provides some insight for choosing N_P.

Effect of reduced measurement set

The performance of a pattern-matching technique is clearly dependent upon the available measurements. The simulations results for the reduced measurement set in Table 3 are summarized in Table 11 for optimum values of the cutoff factors. A comparison of Tables 5 and 11 indicates that the poorer performance for the reduced measurement set is mainly due to the reduction in the pool accuracy, p; however, the pattern-matching efficiency, η, improves slightly. For the S_{PCA}-S_{LV} method, p decreases by almost 36%. The S_{PCA}-S_{LV} method provides the best pattern matching for both the full- and reduced-measurement sets.

A comparison of Tables 5 and 11 indicates that the performance of the T^2, Q, the combined discriminant similarity factors, and S_{LV} methods is less strongly affected by the reduced-measurement set. By contrast, the stronger effect occurs for the S_{PCA} method, because these similarity factors are calculated by matching the directions of the subspaces. When there are fewer directions to compare, the discriminating power of the similarity factor is reduced. As an extreme example, if the measurement set were reduced to only two measurements, then performance would be adversely affected because the rotations of the two-dimensional subspaces may not be sufficient to provide discrimination between many different operating conditions. This consideration is the primary reason why the PCA-based similarity fac-

Table 11. Best Performance for the Reduced-Measurement Set and 5-s Data

Method	Best Cutoff(s)	N_P	p (%)	η (%)	$\eta_{\max}$ (%)	ξ (%)
T^2 statistic (95% limit)	N/A	153	15	80	100	**47**
Q statistic (95% limit)	N/A	103	18	65	100	**41**
Combined Q and T^2	N/A	123	16	95	100	**55**
PCA similarity factor, S_{PCA}	0.981	120	33	93	100	**63**
Limit violation similarity factor, S_{LV}	0.560	101	23	92	100	**58**
PCA and LV similarity factors	**0.981, 0.560**	**84**	**46**	**90**	**100**	**68**

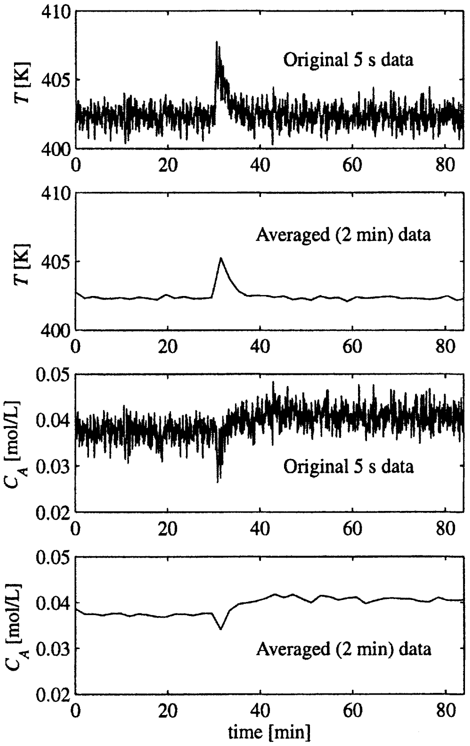

Figure 5. Effect of data averaging for operating condition F6 (step change in Q_F at $t = 30$ min).

tors are more adversely affected by reducing the measurement set.

Effect of averaging process data

The effect of averaging process measurements is shown in Figure 5 for two important process variables: reactor temperature T and concentration C_A. The 5-s sampled data are noisy but preserve the dynamic features of the signal, while 2-min averaging reduces the effect of the step change by suppressing the peak values. Thus, averaging results in loss of dynamic response information. A comparison of Tables 5 and 11–13 indicates that the pattern matching is less successful for the averaged data than for the original data. The reduced performance can be attributed to the loss in dynamic response information due to averaging.

Table 12. Best Performance for the Full-Measurement Set and 2-min Averaged Data

Method	Best Cutoff(s)	N_P	p (%)	η (%)	$\eta_{\max}$ (%)	ξ (%)
T^2 statistic (95% limit)	N/A	78	16	47	100	**31**
Q statistic (95% limit)	N/A	19	6	9	94	**8**
Combined Q and T^2	N/A	36	14	29	99	**22**
PCA similarity factor, S_{PCA}	0.839	85	26	91	100	**58**
Limit violation similarity factor, S_{LV}	0.860	46	49	72	99	**60**
PCA and LV similarity factors	**0.839, 0.860**	**25**	**59**	**65**	**97**	**62**

Table 13. Best Performance for the Reduced-Measurement Set and 2-min Averaged Data

Method	Best Cutoff(s)	N_P	p (%)	η (%)	$\eta_{\max}$ (%)	ξ (%)
T^2 statistic	N/A	111	13	47	100	**30**
Q statistic	N/A	59	12	26	100	**19**
Combined Q and T^2	N/A	71	13	45	100	**29**
PCA similarity factor, S_{PCA}	0.875	197	14	91	100	**53**
Limit violation similarity factor, S_{LV}	0.560	143	20	91	100	**56**
PCA and LV similarity factors	**0.875, 0.560**	**94**	**27**	**83**	**100**	**55**

The S_{PCA}-S_{LV} combination significantly improves the results because the p value increases more than twofold, while there is only a moderate decrease in η. A comparison of the best performing pattern-matching techniques for the 5-s and averaged data is presented in Figure 6, where the S_{LV} cutoff values from Tables 5 and 12 were used. These results show that the proposed pattern-matching methodology is able to match patterns successfully for both the 5-s and the averaged data.

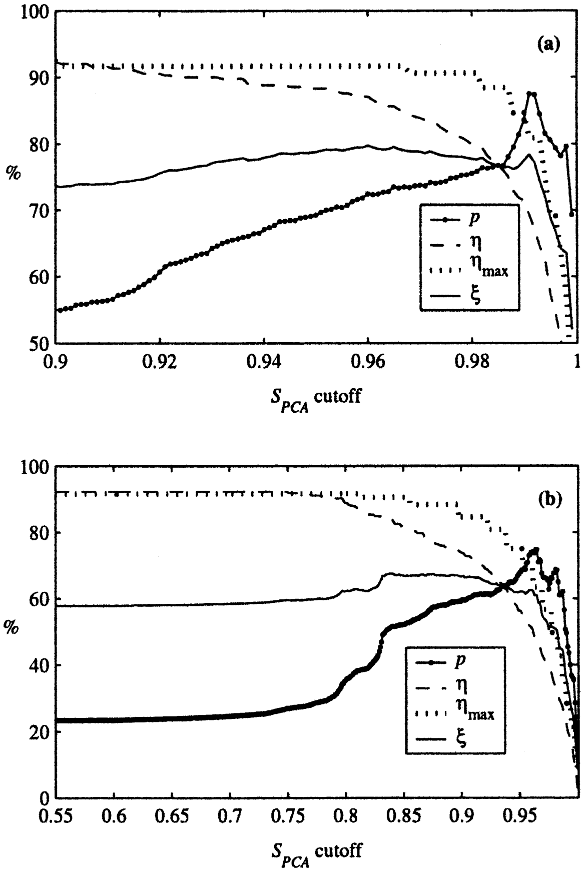

Figure 6. Best results for the full measurement set and S_{PCA}-S_{LV} method.

(a) 5-s data, and (b) 2-min averaged data.

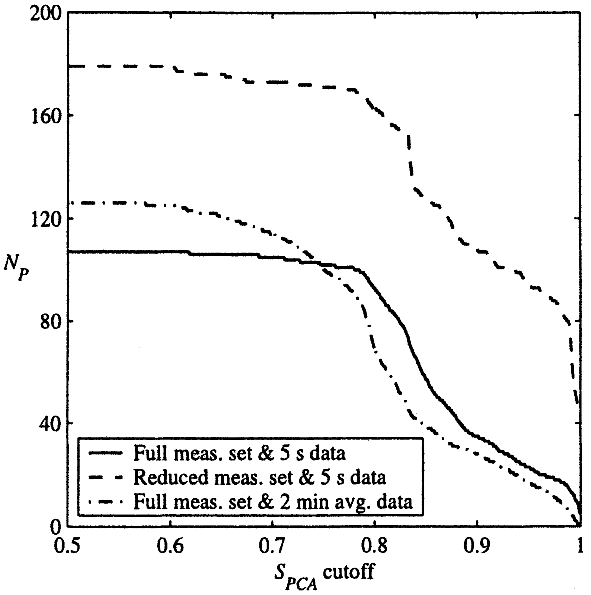

Figure 7. Effect of S_{PCA} cutoff on the candidate pool size, N_P.

Effect of similarity factor cutoff on candidate pool size

Finally, the effect of the similarity factor cutoff on candidate pool size, N_P, is considered to provide insight into the performance of the pattern-matching techniques. In practice, it is not necessary to specify cutoff values in order to generate a candidate pool, as described earlier. Intuitively, as a cutoff value is increased, N_P would be expected to decrease monotonically. This trend is illustrated in Figure 7 for the full measurement set and 5-s data. The cutoffs values for the S_{LV} are the same as for Tables 5, 11, and 12. For very low cutoffs, N_P is quite large, but larger cutoff values reduce N_P to a reasonable size.

Conclusions

A novel strategy has been developed for the diagnosis of abnormal plant operation based on locating previous occurrences in large historical databases. The new pattern-matching strategy is both "data driven" and unsupervised because neither process models nor training data are required. Instead, the pattern matching relies on two similarity factors, including a new similarity factor that characterizes the patterns of alarm limit violations.

The proposed pattern-matching strategy has been evaluated in an extensive simulation study for a continuous stirred-tank reactor system. The historical database for the CSTR included over 474,000 measurements of 14 process variables for 28 different operating modes. The operating modes included a wide variety of faults, disturbances, and process changes. The historical data for both the case studies were generated with the magnitude of the fault (or disturbance or process change) varying randomly between 25 and 125% of its nominal value. The proposed pattern-matching strategy was very effective in locating previous occurrences of a current "abnormal plant operation" in the historical database for both the case studies. In particular, the new strategy was more effective than existing PCA monitoring techniques based on the Q andT^2 statistics and the standard PCA similarity factor alone.

Acknowledgments

Financial support from the UCSB Process Control Consortium, Pavilion Technologies, Inc., Chevron Research & Technology Co., and OSI Software, Inc. is gratefully acknowledged.

Notation

A = cross-sectional area of the reactor, dm^2
A_C = area available for heat transfer, dm^2
C_A = concentration of species A in reactor, mol/L
C_{AF} = concentration of species A in reactor feed stream, mol/L
C_p = heat capacity of reactor contents, J/g·K
C_{pC} = heat capacity of coolant, J/g·K
C_v = valve coefficient, L/min·psi$^{1/2}$
E = activation energy, J/mol
$f(l)$ = valve characteristic
g_s = specific gravity of the fluid
h = liquid level in the reactor, dm
ΔH = heat of reaction, J/mol
k_0 = preexponential factor, min^{-1}
l = fraction that the valve is open (0 = closed; 1 = fully open)
LV_c = critical number of high-limit or low-limit violations
N_1 = the number of records in the candidate pool that are actually similar to the current snapshot
N_2 = the number of records in the candidate pool that are not similar to the current snapshot
N_{DB} = the total number of historical data windows that are similar to the current snapshot
N_P = candidate pool size = $N_1 + N_2$
p = pool accuracy (%) = $(N_1/N_p)\times 100\%$
ΔP_v = pressure drop across the valve, psi
Q = flow rate of reactor outlet stream, L/min
Q_C = coolant flow rate, L/min
Q_F = feed flow rate of reactor feed stream, L/min
R = universal gas constant, J/mol·K
T = temperature in reactor, K
T_C = temperature of coolant in the cooling jacket, K
T_{CF} = temperature of coolant feed, K
T_F = temperature of reactor feed stream, K
U = heat-transfer coefficient, J/min·K·dm^2

Greek letters

η = pattern matching efficiency (%) = $(N_1/N_{DB})\times 100\%$
ρ = density of reactor contents, g/L
ρ_C = density of coolant, g/L
ξ = average of pool accuracy and pattern matching efficiency (%) = $(p+\eta)/2$

Literature Cited

Agrawal, R., K. Lin, H. S. Sawhney, and K. Shim, "Fast Similarity Search in the Presence of Noise, Scaling and Translation in Time Series Databases," *Proc. Int. VLDB Conf.*, Zurich, Switzerland, p. 490 (1995).

Anonymous, Exhibit at the *Edmonton Space and Science Centre*, Edmonton, Canada (2000).

Apté, C., "Data Mining: An Industrial Research Perspective," *IEEE Trans. Comput. Sci. Eng.*, **4**, 6 (1997).

Bishop, C. M., *Neural Networks for Pattern Recognition*, Claredon Press, Oxford (1995).

Chiang, L. H., E. L. Russel, and R. D. Braatz, *Fault Detection and Diagnosis in Industrial Systems*, Springer-Verlag, London (2001).

Davis, J. F., M. J. Piovoso, K. A. Kosanovich, and B. R. Bakshi, "Process Data Analysis and Interpretation," *Advances in Chemical*

Engineering, Vol. 25, J. Wei, J. L. Anderson, K. Bischoff, and J. Seinfeld, eds., Academic Press, New York, p. 1 (1999).

Dehaspe, L., H. Toivonen, and R. C. King, "Finding Frequent Substructures in Chemical Compounds," *Proc. Int. Conf. on Knowledge Discovery and Data Mining*, New York (1998).

Dempster, A. P., N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data Via the EM Algorithm," *J. Roy. Stat. Soc., Ser. B*, **39**, 1 (1977).

Duda, R. O., and P. E. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York (1973).

Duda, R. O., P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed., Wiley, New York (2001).

Faloutsos, C., M. Ranganathan, and Y. Manolopoulos, "Fast Subsequence Matching in Time Series Databases," *Proc. ACM SIGMOD-94*, Minneapolis, MN, p. 419 (1994).

Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth, "The KDD Process for Extracting Useful Knowledge from Volumes of Data," *Commun. ACM*, **39**, 27 (1996).

Fukunaga, K., *Introduction to Statistical Pattern Recognition*, Academic Press, New York (1990).

Gavrilov, M., D. Anguelov, P. Indyk, and R. Motwani, "Mining the Stock Market: Which Measure is Best?" *Proc. ACM SIGKDD Conf. on Knowledge Discovery and Data Mining*, Boston, MA, p. 487 (2000).

Jackson, J. E., *A User's Guide to Principal Components*, Wiley, New York (1991).

Johannesmeyer, M. C., *Abnormal Situation Analysis Using Pattern Recognition Techniques and Historical Data*, MSc Thesis, Univ. of California, Santa Barbara (1999).

Jolliffe, I. T., *Principal Component Analysis*, Springer-Verlag, New York (1986).

Kavuri, S. N., and V. Venkatasubramanian, "Representing Bounded Fault Classes Using Neural Networks With Ellipsoidal Activation Functions," *Comput. Chem. Eng.*, **17**, 139 (1993).

Kennedy, R. L., Y. Lee, B. Van Roy, C. D. Reed, and R. P. Lippman, *Solving Data Mining Problems Through Pattern Recognition*, Prentice Hall, Englewood Cliffs, NJ (1998).

Keogh, E., K. Chakrabarti, M. Pazzani, and S. Mehrotra, "Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases," *Knowl. Inf. Syst.*, **3**, 263(2001).

Koivo, H., "Artificial Neural Networks in Fault Diagnosis and Control," *Control Eng. Pract.*, **2**, 89 (1994).

Kourti, T., and J. F. MacGregor, "Multivariate SPC Methods for Process and Product Monitoring," *J. Qual. Technol.*, **28**, 409 (1996).

Kourti, T. J., J. Lee, and J. F. MacGregor, "Experiences with Industrial Applications of Projection Methods for Multivariate Statistical Process Control," *Comput. Chem. Eng.*, **20**, S745 (1996).

Kramer, M. A., and R. S. H. Mah, "Model Based Monitoring," *Proc. Int. Conf. on Foundations of Computer Aided Process Operations*, CACHE, Austin, TX, p. 45 (1994).

Krzanowski, W. J., "Between-Groups Comparison of Principal Components," *J. Amer. Stat. Assoc.*, **74**, 703 (1979).

Ku, W., R. H. Storer, and C. Georgakis, "Disturbance Detection and Isolation by Dynamic Principal Component Analysis," *Chemometrics Intell. Lab. Syst.*, **30**, 179 (1995).

Martin, E. B., and A. J. Morris, "An Overview of Multivariate Statistical Process Control in Continuous and Batch Performance Monitoring," *Trans. Inst. Meas. Control*, **18**, 51 (1996).

Mylaraswamy, D., "Experiences From Fielding an Abnormal Event Guidance System," *Preprints of IFAC Workshop on On-Line Fault Detection and Supervision in the Chemical Process Industry*, Jejudo Island, Korea, p. 277 (2001).

Perng, C.-S., H. Wang, S. R. Zhang, and D. S. Parker, "Landmarks: A New Model for Similarity-Based Pattern Querying in Time Series Databases," *Int. Conf. on Data Engineering (ICDE' 2000)*, San Diego, CA (2000).

Raich, A., and A. Çinar, "Statistical Process Monitoring and Disturbance Isolation in Multivariate Continuous Processes," *Proc. IFAC-ADCHEM'94*, Kyoto, Japan, p. 452 (1994).

Raich, A., and A. Çinar, "Multivariate Statistical Methods for Monitoring Continuous Processes: Assessment of Discrimination Power of Disturbance Models and Diagnosis of Multiple Disturbances," *Chemometics Intell. Lab. Syst.*, **30**, 37 (1995).

Raich, A., and A. Çinar, "Statistical Process Monitoring and Disturbance Diagnosis in Multivariate Continuous Processes," *AIChE J.*, **42**, 995 (1996).

Raich, A., and A. Çinar, "Diagnosis of Process Disturbances by Statistical Distance and Angle Measures," *Comput. Chem. Eng.*, **21**, 661 (1997).

Ramakrishnan, R., S. Stolfo, R. Bayardo, and I. Parsa, eds., *Proc. ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, Boston, MA (2000).

Regaldo, A., "Mining the Genome," *Technol. Rev.*, **102**(6), 56 (1999).

Russo, L. P., and B. W. Bequette, "Effect of Process Design on the Open-Loop Behavior of a Jacketed Exothermic CSTR," *Comput. Chem. Eng.*, **20**, 417 (1996).

Seborg, D. E., T. F. Edgar, and D. A. Mellichamp, *Process Dynamics and Control*, Wiley, New York (1989).

Shürmann, J., *Pattern Classification: A Unified View of Statistical and Neural Approaches*, Wiley, New York (1996).

Singhal, A., *Pattern Matching in Multivariate Time-Series Data*, PhD Thesis, Univ. of California, Santa Barbara (2002).

Singhal, A., and D. E. Seborg, "Dynamic Data Rectification Using the Expectation Maximization Algorithm," *AIChE J.*, **46**, 1556 (2000).

Smyth, P., "Probabilistic Model-Based Clustering of Multivariate and Sequential Data," *Proc. Int. Workshop on AI and Statistics*, D. Heckermann and J. Whittaker, eds., Los Gatos, CA (1999).

Sorsa, T., and H. Koivo, "Application of Artificial Neural Networks in Process Fault Diagnosis," *Automatica*, **29**, 843 (1993).

Stephanopoulos, G., and C. Han, "Intelligent Systems in Process Engineering: A Review," *Proc. Int. Symp. on Process Systems Engineering (PSE 94)*, Kyongju, Korea, p. 1339 (1994).

Vaidyanathan, R., and V. Venkatasubramanian, "Representing and Diagnosing Dynamic Process Data Using Neural Networks," *Eng. Appl. Artif. Intell.*, **5**, 11 (1992).

Valle, S., W. Li, and S. J. Qin, "Selection of the Number of Principal Components: The Variance of the Reconstruction Error Criterion with a Comparison to Other Methods," *Ind. Eng. Chem. Res.*, **38**, 4389 (1999).

Vedam, H., and V. Venkatasubramanian, "PCA-SDG Based Process Monitoring and Fault Diagnosis," *Control Eng. Pract.*, **26**, 903 (1999).

Wang, X. Z., *Data Mining and Knowledge Discovery for Process Montoring and Control*, Springer-Verlag, London (1999).

Wang, X. Z., "Knowledge Discovery Through Mining Process Operational Data," *Application of Neural Networks and Other Learning Technologies in Process Engineering*, I. M. Mujtaba and M. A. Hussain, eds., Chap. 13, Imperial College Press, London, p. 287 (2001).

Wang, X. Z., and C. McGreavy, "Automatic Classification for Mining Process Operational Data," *Ind. Eng. Chem. Res.*, **37**, 2215 (1998).

Whiteley, J. R., and J. F. Davis, "A Similarity-Based Approach to Interpretation of Sensor Data Using Adaptive-Resonance Theory," *Comput. Chem. Eng.*, **18**, 637 (1994).

Wise, B. M., and N. B. Gallagher, *PLS Toolbox 2.1: User's Manual*, Eigenvector Research, Inc., Manson, WA (2000).

Zhang, J., E. Martin, and A. J. Morris, "Fault Detection and Classification Through Multivariate Statistical Techniques," *Proc. Amer. Control Conf.*, Washington, DC, p. 751 (1995).

Manuscript received Dec. 4, 2000, and revision received Feb. 15, 2002.