Fault Diagnosis of Chemical Processes with Incomplete Observations: A Comparative Study

M. Askarian[a1], G. Escudero[b], M. Graells[c], R. Zarghami[a], F. Jalali-Farahani[a], N. Mostoufi[a]*

[a] School of Chemical Engineering, College of Engineering, University of Tehran, PO Box 11155-4563, Tehran, Iran.

[b] Computer Science Department. Universitat Politècnica de Catalunya. EUETIB, Comte d'Urgell 187, 08028-Barcelona, Spain.

[c] Chemical Engineering Department. Universitat Politècnica de Catalunya. EUETIB, Comte d'Urgell 187, 08028-Barcelona, Spain.

ABSTRACT

An important problem to be addressed by diagnostic systems in industrial applications is the estimation of faults with incomplete observations. This work discusses different approaches for handling missing data, and performance of data-driven fault diagnosis schemes. An exploiting classifier and combined methods were assessed in Tennessee-Eastman process, for which diverse incomplete observations were produced. The use of several indicators revealed the trade-off between performances of the different schemes. Support vector machines (SVM) and C4.5, combined with *k*-nearest neighbourhood (*k*NN), produce the highest robustness and accuracy, respectively. Bayesian networks (BN) and centroid appear as inappropriate options in terms of accuracy, while Gaussian naïve Bayes (GNB) is sensitive to imputation values. In addition, feature selection was explored for further performance enhancement, and the proposed contribution index showed promising results. Finally, an industrial case was studied to assess informative level of incomplete data in terms of the redundancy ratio and generalize the discussion.

KEYWORDS: Fault diagnosis, Missing data, Incomplete observations, Classification, Imputation, Machine learning.

[1]Chemical Engineering Department. Universitat Politècnica de Catalunya. EUETIB, Comte d'Urgell 187, 08028-Barcelona, Spain.

* Corresponding author. Tel.: +34 93 413 72 75.E-mail address: moises.graells@upc.edu

1. INTRODUCTION

Due to the increasing complexity of modern industrial processes, preventive monitoring and fault diagnosis (FD) have become essential to ensure safe operation, improve product quality and sustain economic profit of manufacturing. A new generation of digital instruments, data processing devices, automation systems and high-performance computing tools, have promoted the development of a smart platform for plant monitoring and diagnosis. In addition, various methodologies have been developed to address the FD challenge (Venkatasubramanian et al., 2003). Among them, data-driven diagnosis methods offer appropriate solutions to many difficulties arising in this field (MacGregor and Cinar, 2012; Qin, 2012; Yin et al., 2012). This work investigates the performance of different data-driven FD approaches submitted to increasing loss of data.

FD is mainly a classification problem and machine learning provides various tools for classification. Based on the decision boundaries (hypothesis space source), classification approaches have 4 major types as listed below (Marquez et al., 2007):

- **Distance based methods**. e.g. *k*-nearest neighbourhood (*k*NN) (Duda et al., 2001) and centroid (Salton, 1989);
- **Rule based methods**. e.g. C4.5, as an extension of decision tree (Quinlan, 1993), and AdaBoost (Schapire and Singer, 1999);
- **Probabilistic methods**. e.g. Bayesian network (BN) (Pearl, 1988) and Gaussian naïve Bayes (GNB) (Friedman et al., 1997);
- **Margin based methods**. e.g. support vector machines (SVM) (Cristianini and Taylor, 2000) and artificial neural networks (Duda et al., 2001).

Performance of classifiers is affected by the quality of data, and there are potential factors that may cause incomplete data sets. Figure 1 shows the typical flow of data in a chemical plant and potential causes of missing data. In fact, chemical systems include different blocks, such as process, control, data acquisition, FD systems and human interface. Generally, sensors readings, off-line analyses, and records of actuators status constitute the input data source of the FD system in a chemical plant. These data are also called features or observations in the classification literature and they are used indistinctly. Furthermore, there are various types of media for transmitting data, such as wire, wireless (satellite and radio) and fiber-optic. In any case, different hardware, software or administrative problems

are likely to disrupt access of the FD system to complete datasets. Sensors and actuators are subject to different occasional failures due to wear, abrasion, exposure to physical damage or lack of energy supply (Kadlec et al., 2009; Zhao and Fu, 2015). They may be out of service due to maintenance or removal. In addition, improper calibration of sensors or control valves leads to unreliable or out of range data (Wang et al., 2007). Thus, the records may be automatically deleted by the preprocessing block of the monitoring system (e.g. outlier detection). Furthermore, different types of measurements and sampling rates (e.g. off-line analysis vs. on-line sensing) may produce sparse datasets (Kadlec et al., 2009) and information may be delayed by administrative issues (Warne et al., 2004). Moreover, temporary unavailability of data may happen due to malfunction or disruption of the data acquisition system, which have various causes such as higher bit rate error, short circuit of cables, break of circuit, induced current or even harsh weather (ISO, 1994; Ji and Elwalid, 2002; Rodriguez et al., 2002).

Based on the above discussion, in an industrial practice it is necessary to deal with incomplete datasets and unknown measurements, while continuously demanding useful and reliable information to support decision-making. Complexity of this issue depends on the mechanism of missing data and the informative level of the process database. Three standard mechanisms are usually considered (Rubin, 1976): missing completely at random (MCAR), missing at random (MAR), not missing at random (NMAR). MCAR and MAR are called ignorable mechanisms, and it is not required to take into account causes of missing data in the analysis of the datasets (Schafer, 1997). However, the NMAR mechanism leads to loss of valuable information that cannot be properly compensated (García-Laencina et al., 2010). On the other hand, the informative level of a database depends on the existence of co-linear variables, partial redundancy or duplication of data, which can simplify the analysis of missing data (Kadlec et al., 2009). Determining the informative level of a dataset was studied by Auffarth et al. (2010) and Peng et al. (2005)

Considering the above-mentioned facts, the performance of different FD methodologies to cope with incomplete input datasets needs to be examined. Since most diagnostic systems cannot assign a fault to incomplete features, it is required to consider additional procedures along with conventional classifiers. Indeed, FD with missing data concerns two different problems, handling missing values and classification. As shown in Figure 2, reported

approaches for solving these problems can be grouped into four types (García-Laencina et al., 2010; Scheffer, 2002; Sharpe and Solly, 1995)

- Deletion of the incomplete feature vectors, and classification of the complete data portion only;
- Development of a multi-classifier corresponding to all combination of feature subsets, and classification of incomplete data using the model trained by the same available features;
- Imputation or estimation of missing data, and classification using the edited set;
- Implementation of exploiting procedures for which classification can be still accomplished in the presence of missing variables.

In the first and third approaches, handling missing values (deletion and imputation, respectively) and classification are two problems that should be solved separately. In contrast, the second and fourth approaches are able to directly handle incomplete input datasets (Ahmad and Tresp, 1993; Huang, 2008). The first alternative is too simplistic and may be unacceptable in many applications. The second approach (multi-classifier) requires a set of training models based on all possible combinations of features. Sharpe and Solly (1995) have reported this promising approach to be more efficient than others while dealing with very limited number of faults and features. Although numerous features exist in chemical processes, this approach quickly explodes in terms of complexity and number of classifier models (Gabrys, 2002). Therefore, this work addresses and discusses the last two approaches (grey boxes in Fig. 2).

Estimation of missing values of a dataset, which is a preliminary step in the third approach, can be accomplished via various methodologies (Fortuna et al., 2007; Gonzalez, 1999; Little and Rubin, 2014). Generally, they can be grouped into four types in spite of the overlap between some basic principles:

- **Regression:** e.g. least squares (Chen and Chen, 2000), auto-regression and moving-average models (Palit and Popovic, 2006) as well as *k*NN (Batista and Monard, 2002);

- **Statistical method:** e.g. mean (García-Laencina et al., 2010), maximum-likelihood (ML) and expectation–maximization algorithms (EM) (Ibrahim, 1990; Walczak and Massart, 2001b), principal component analysis (PCA) and partial least squares (PLS) (Kadlec et al., 2011; Nelson et al., 2006; Walczak and Massart, 2001a);
- **Soft computing**: e.g. artificial neural network (ANN) (Abdella and Marwala, 2005; Bishop, 1995) and fuzzy inference system (FIS) (Atkeson et al., 1997; Luo and Shao, 2006);
- **Phenomenological models:** e.g. first principle models such as energy and mass balance (Chéruy, 1997) and adaptive observers (Bastin and Dochain, 1990).

Moreover, attention has also been paid to hybrid methods (Kadlec et al., 2009). However, phenomenological models require deep understanding of the underlying physical and chemical phenomena, and are usually limited to ideal steady states. On the other hand, faults often lead to transient and/or unsteady state in chemical processes. Therefore, this work covers data-driven imputation models which readily reflect real conditions of process systems (grey boxes in Fig. 2).

Nelson et al. (1996) studied process monitoring in the presence of missing measurements using principal component analysis (PCA) and partial least squares (PLS). They presented three approaches for estimating scores: a single component projection method, a method of simultaneous projection to the model plane, and conditional mean imputation. Their analysis of a Kamyr pulp digester with 22 measurements and up to 20% missing data revealed that the conditional mean replacement method is generally superior to the other approaches. Subsequently, uncertainty intervals were derived for assessing the performance of monitoring applications with incomplete observations. The size of uncertainty regions helped to distinguish between situations where model performance with missing measurements was acceptable or not (Nelson et al., 2006).

The last approach of FD with missing data is exploiting procedures that have robust algorithms when coping with missing features. They can keep on classifying even in the presence of incomplete input data. For example, ID3 is an extension of the decision tree algorithm that handles an unknown feature by generating an additional edge for the unknown value. Thus, the classification of an incomplete dataset is fulfilled by taking into

account an unknown edge like other values (Sharpe and Solly, 1995). Nevertheless, the results have revealed that ANN is superior to ID3 in terms of classification performance. Furthermore, Huang (2008) proposed to take advantage of BN and the fact that it needs no change in the model and the network setup to cope with incomplete datasets. However, the reported results cannot be generalized to other cases due to generating redundant information by feature extension in that work. In other words, robustness of the BN was assessed in terms of missing artificial features rather than missing original data. Thus, a more comprehensive framework for evaluation of performance is required.

The aim of the above mentioned works was producing reliable diagnosis from new incomplete observations. However, developing an FD system from incomplete training data has also attracted researchers' attention. In this regard, various algorithms have been developed, including Bayesian PCA (BPCA) (Ge and Song, 2011), Hopfield neural networks (Wang, 2005) and training-estimation-training (TEST) (Yoon and Lee, 1999). This issue is important for developing a model with small sample size or online updating of the model. Nevertheless, since there exists a large amount of historical data in chemical plants, it is possible to provide a complete training dataset, because observations with missing values can be discarded from the training subset based on the assumption that observations are independent. Thus, designing a diagnostic system for chemical plants with training data contained missing measurements may not be a critical issue. Hence, this work is focused on the FD with new incomplete observations.

FD of chemical processes with missing data has not been satisfactorily addressed in the literature. Therefore, there is a need to provide an appropriate FD framework dealing with incomplete data. Comprehensive assessment of various tools, as well as identification of their advantages and limitations is required. In the present work, standard centroid, C4.5, GNB, and SVM methods, as representative of each type of classification approaches (Marquez et al., 2007), were investigated while missing values were artificially produced in the datasets. Diagnosis performance of the exploiting procedure (BN), facing incomplete data, was specified. Imputation was accomplished using mean, PCA, *k*NN and ANN as statistical, regression, and soft computing methods, respectively. Combinations of different imputation and classification models were examined. Various indexes, such as accuracy, robustness and sensitivity, were defined for an evaluation purpose. Moreover, enhancing

the performance of the classification was explored through feature extraction. Finally, a new index was proposed to determine the level of redundancy, which facilitates handling of incomplete data.

2. METHODOLOGY

2.1. Tools

In this section, classification and imputation methods used in the subsequent sections are introduced.

2.1.1. Classification

Different types of classifiers are listed in Section 1. In the present work, FD systems are developed by following methods:

- **Centroid** (distance based), which is based on "mean difference" as a simple and intuitive concept of discrimination (Jiang et al., 2009).
- **C4.5** (rule based), which builds decision trees consisting of sets of ordered rules based on information entropy (Amarnath et al., 2013). Then, the decision tree that best matches a new observation in terms of highest weight reflects the type of fault.
- **GNB and BN** (probabilistic), which consist of maximizing the conditional probability of a particular fault given a set of observations (Liu et al., 2010; Verron et al., 2007). A Gaussian probability and a conditional probability table (CPT) are assigned to each variable using GNB and BN, respectively.
- **SVM** (margin based), which discriminates different fault instances, as positive and negative points, using a kernel function (hyperplane) that maximizes the margin (Yélamos et al., 2009).

2.1.2. Imputation

Most monitoring or diagnosis algorithms fail in case of missing data, and imputation is suggested as an alternative solution (Fig. 2). In this study, four imputation options are examined to edit an incomplete dataset: the mean imputation and PCA, which are statistical methods, the *k*NN imputation, which is a regression, and ANN, which is a soft computing method.

The mean imputation is the earliest and easiest approach that fills the missing point by the average value of that variable in the historical records. Another statistical alternative is an iterative algorithm based on PCA. It consists of initial estimation of missing values, decomposition to principal components, reconstruction of features, and replacing of prediction in missing variables until convergence (Walczak and Massart, 2001a). Multi-layer perceptron (MLP), which is one of the most popular ANN-based models, utilizes supervised learning technique, called back-propagation. Thereafter, it is used for estimation of missing values based on other available values. The *k*NN method is a distance based method, in which *k* (number of nearest neighbours) donors are selected from the complete historical database, so that they minimize a distance function. Then, the average of corresponding variable in the *k* donors is replaced in the missed point.

It is worth noting that although the *k*NN method is usually implemented for classification purposes, in the present work it is used for regression. Furthermore, it is more advantageous to apply the *k*NN method rather than conventional regression methods, in which explicit model should be trained. Based on the ¨lazy learning¨ approach of *k*NN that locally approximates the distance function, no training model construction is required (Zhang and Zhou, 2007). In this way, the *k*NN method can be easily applied even if various measurements are missed at each time step.

2.2. Design of Experiments

Figure 3 illustrates the general procedure for evaluation of FD performance whenever datasets are incomplete. The procedure is described as follows:

Step 1: The original data are arranged in a matrix in which each row represents the time series of a measured variable. For each state of the system (faulty and normal), T^f samples of V variables are recorded in the matrix:

$$\mathbf{X}^f = \begin{bmatrix} x_{11}^f & x_{12}^f & \cdots \\ x_{21}^f & x_{22}^f & \cdots \\ \vdots & \vdots & \ddots \\ x_{V1}^f & x_{V2}^f & \cdots \end{bmatrix} \quad \vdots \qquad \mathbf{X}^f \in R^{V \times T}; \quad v=1, 2, \ldots, V \quad t=1, 2, \ldots, T^f \qquad (1)$$

where x_{vt}^f is the value of the v^{th} measured variable at the sampling time t in the f^{th} state of the system. Then, the collection of F various states are considered as a labelled process dataset:

$$\mathbf{X\!L} = \left\{ \mathbf{X}^f \mid f = 1,2,...,F \right\} \quad \mathbf{X\!L} \in R^{F \times V \times T^f} \tag{2}$$

Step 2: The labelled process dataset, $\mathbf{X\!L}$, is split into training matrix, $\mathbf{XL}$, and a test matrix, $\mathbf{X}$, which are used for fitting the parameters of the models and estimating of unknown faults, respectively. For the sake of simplicity, in the subsequent sections, the f^{th} index of measurements is ignored (e.g. x_{vt}) whenever dealing with the test dataset.

Step 3: The training matrix, $\mathbf{XL}$, is standardised as follows, to make the algorithm less sensitive to particular variables:

$$\frac{\mu_v}{\sigma_v} \tag{3}$$

$$\mu_v = \frac{\sum_{f=1}^{F} \sum_{t=1}^{\tilde{T}} x_{vt}^f}{\sum_{f=1}^{F} \tilde{T}} \tag{4}$$

$$\sigma_v = \sqrt{\frac{\sum_{f=1}^{F} \sum_{t=1}^{\tilde{T}} (x_{vt}^f - \mu_v)^2}{\sum_{f=1}^{F} \tilde{T}}} \tag{5}$$

where $\tilde{T}$ is the number of training samples for each fault.

Step 4: Parameters of each classifier, mentioned in section 2.1, are fitted based on the standardised training matrix, $\dot{\mathbf{X}}\mathbf{L}$, obtained in the previous step.

Step 5: Some measurements are artificially deleted from the complete test dataset, $\mathbf{X}$. In particular, 4 incomplete test matrices were produced in this work by random deletion of 10%, 20%, 30% and 40% of measurements. All test matrices are scaled using mean (Eq.4) and variance (Eq.5) of the training dataset.

Step 6: Imputation methods are applied to configure an estimated full test dataset, $\dot{\mathbf{X}}$, based on the complete training dataset, $\dot{\mathbf{X}}\mathbf{L}$. The BN, as an exploiting method, can skip the imputation step to predict faults with incomplete data.

Step 7: The efficiency of different imputation methods is evaluated by the Pearson correlation as the predictive accuracy (PAC) index:

$$PAC = \frac{\sum_{n=1}^{N} (\quad \cdot \quad \cdot \quad \cdot}{\sqrt{\sum_{n=1}^{N} (\quad \cdot \quad \cdot \quad \cdot \quad \sqrt{N}}} \qquad (6)$$

where N is the number of missing values; $\quad$ and $\quad$ are true (in $\mathbf{X}$) and imputed (in $\mathbf{X}$) values of the v^{th} variable, respectively; and $\quad$ and $\quad$ are the means of the true and imputed values of the v^{th} variable, respectively. A good imputation method will produce a PAC value close to 1.

Step 8: The snapshot of the test dataset at each time step, $[\quad$, is assigned to the fault that is predicted by the classifier.

Step 9: The performance of classifiers is evaluated by comparing predicted faults and true faults for each snapshot. In this order FD outcomes are arranged in a confusion matrix as presented in Table 1 (Yélamos et al., 2007). Then, performance indexes including accuracy, precision, recall, and *F1* are calculated as follows:

$$Accuracy = \frac{a+d}{a+b+c+d} \qquad (7)$$

$$Precision = \frac{a}{a+b} \qquad (8)$$

$$Recall = \frac{a}{a+c} \qquad (9)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \qquad (10)$$

where a is the number of samples corresponding to faulty situations and diagnosed as such (true positive); b is the number of samples diagnosed as faulty but were not (false positive); c is the number of samples corresponding to faulty but not diagnosed situations (false negative) and d is the number of samples not happened and not diagnosed (true negative). The classifier performance regarding each individual fault is assessed through *F1,* which can be interpreted as a weighted average of the precision and recall. The accuracy index is appropriate for global evaluation.

3. TEP CASE STUDY

The Tennessee-Eastman process (TEP) proposed by Downs and Vogel (1993) is widely used for testing diagnosis techniques. The dynamic simulation of this process includes a

control strategy created by Ricker (1996). Figure 4 illustrates the process and instrumentation diagram of the TEP, which consists of five major unit operations, including a reactor, a product condenser, a vapour–liquid separator, a recycle compressor and a product gas stripper. Two products (*G* and *H)* are produced by two simultaneous gas–liquid exothermic reactions and a byproduct (*F)* is generated by two additional exothermic reactions from reactants *A, C, D* and *E*. This process has 12 manipulated variables, 22 continuous measurements and 19 composition measurements. The 20 pre-defined faults of the TEP are presented in Table 2. The TEP includes random error and white noises that impact on diagnosis complexity.

The original datasets, $\mathbf{X}^f$, were generated by the simulation of a 30-h operation horizon for each fault and the normal situation. Sampling every 0.1 h resulted in a labelled process data matrix, $\mathbf{X_L} \in R^{21\times53\times300}$. Training dataset, $\mathbf{X_L} \in R^{21\times53\times180}$, and test dataset, $\mathbf{X} \in R^{21\times53\times120}$, were built by randomly splitting samples (60:40 ratio). After standardising the training dataset, the 5 FD systems, centroid, C4.5, BN, GNB, and SVM (default radial based function) were used and assessed. In order to compare their raw performance in case of incomplete observations no parameter tuning was attempted to the standard methods.

4. RESULTS AND DISCUSSION OF THE TEP CASE STUDY

4.1. Performance

As discussed in section 1, the BN is able to exploit all available information, complete or incomplete, to estimate the state of the system (Huang, 2008). In the first stage, the capability of this FD approach was evaluated on the TEP to produce a reference result for subsequent comparison. The performance index, *F1*, for each fault is given in Appendix A. Figure 5 illustrates how the global accuracy of the BN degrades with missing data. The global accuracy of the BN drops below 0.10, while more than 20% of data are not available. On the other hand, simple mean imputation can recover accuracy of the BN to some extent. Furthermore, pre-treatment of the incomplete test data with *3*NN imputation shows significant performance improvement of the BN.

From algorithmic point of view, the BN can keep inference of the state of the system even with incomplete data, as mentioned before. Despite its apparent robustness, the BN classifier has performed poorly in the presence of missing data. The main reason is

existence of numerous faults in the TEP case study that burden the classification. However, BN performance was enhanced when combined with imputation (Fig. 5). When partial data are missing, most methodologies require imputation to edit the test dataset before classification. Subsequently, the performance of centroid, C4.5, BN, GNB and SVM, coupled with different imputation methods, mean, PCA, ANN and *3*NN, were analysed. Figure 6 illustrates the global accuracy of these methodologies as a function of data incompleteness. FD methods can be ranked in terms of global accuracy as follows: C4.5, GNB, SVM, BN and centroid when data are complete (0% missing data). This can be taken as a reference for further comparative purposes. In addition, despite the type of classifier, *3*NN and ANN imputation keeps the performance higher than other imputations due to higher accuracy in terms of PAC (Appendix A).

The accuracy loss due to missing data also depends on the classification along with the imputation choice. In order to quantify this effect, the robustness index is introduced as:

$$Robustness = \frac{Accuracy^{mis}}{Accuracy^{comp}} \tag{11}$$

where $Accuracy^{comp}$ and $Accuracy^{mis}$ are the accuracies obtained by each method when missing data in the test sets, $\mathbf{X}$, is 0% and 40% , respectively. Hence, Figure 7 compares the fault diagnosis methods in terms of robustness. Generally, *3*NN imputation is shown to provide the highest robustness, regardless of the type of classifier. Despite high accuracy of C4.5 in case of complete observations (Fig. 6), it does not keep this behaviour whenever faced with incomplete observations (Fig. 7). In other words, C4.5 is not robust respect to missing data.

Furthermore, the imputation approaches, mean, PCA, ANN and *3*NN, have different impact on the accuracy. The sensitivity of each integrated methodology can be assessed by the following index:

$$Sensitivity = Accuracy^{\max} - Accuracy^{\min} \tag{12}$$

where $Accuracy^{max}$ and $Accuracy^{min}$ are maximum and minimum accuracies obtained by the corresponding classifier when the test sets suffer 40% missing data. Sensitivities of different approaches are compared in Figure 8. This figure shows that SVM and centroid methods are less sensitive than other algorithms.

This behaviour can be explained by the principles (space source hypothesis) of these methods. In the training stage, decision boundaries that can discriminate faults are determined. Training of SVM and centroid includes a loop that determines the optimal boundary in terms of maximum margin. However, training of other methods terminates once a decision boundary is determined. The optimal decision boundary of SVM and centroid allows to better discriminate the feature space regardless of the imputation method. In other words, the widest margin surrounding the boundary increases the tolerance limit of the feature space. This is in accordance with the fact that learning bias has proved to have good properties in terms of generalization bounds for the classifier (Marquez et al., 2007).

4.2. Computational time

The computer configuration used in this work was a core i7-3770@3.40GHz processor with 8GB RAM. Computational times of all models, considering various ratios of missing data, are given in the Appendix A. In general, results reveal that CPU times for classification models are negligible in comparison with imputation models. In terms of increasing CPU time, imputation methods rank as follows: mean, PCA, *3*NN, ANN. On the other hand, *3*NN and ANN have higher accuracy in terms of PAC compared with mean and PCA. Furthermore, Figure 7 shows that integration of any classifiers with *3*NN imputation guarantees the highest robustness of the fault diagnosis. However, *3*NN and ANN require higher computational effort than mean and PCA imputation.

The performance of imputation would be greatly improved if the computational burden can be reduced, and reduction of the search space seems worthy of being explored towards this end. In other words, decreasing the number of variables can help finding the nearest neighbour sooner. Therefore, the subset of measurements that are highly affected by different faults should be selected as the significant feature set. The contribution index ($\mathbf{CI}$) is introduced in this work for the feature selection as:

$$CI_v^f = 1 - \frac{\sum_{i=1}^{F}\sum_{t=1}^{\tilde{T}} \| x_{vt}^f - x_{vt}^i \|}{\sum_{i=1}^{F}\sum_{j=1}^{F}\sum_{t=1}^{\tilde{T}} \| x_{vt}^j - x_{vt}^i \|} \qquad \forall f, \forall v \quad \mathbf{CI} \in R^{F \times V} \tag{13}$$

The deviation of the measurements corresponding to each fault at each time step is assessed by the Euclidean distance. Then, a subset of variables, $\mathbf{V}^f$, consisting of the S elements having the highest CI_v^f is selected as the collection of significant features for each fault:

$$I = \{1, 2, ..., V\} \tag{14}$$

$$\mathbf{V}^f = \left\{ A^f \subset I \;\middle|\; |A^f| = S, \min_{i \in A^f} \{CI_i^f\} \geq \max_{j \in I \setminus A^f} \{CI_j^f\} \right\} \quad \forall f \tag{15}$$

A set of significant features, $\mathbf{V}$, is the union of these subsets:

$$\mathbf{V} = \bigcup \mathbf{V} \tag{16}$$

In the TEP case study, the set $\mathbf{V}$ includes 41 measurements while S is 15. Therefore, variables of the significant feature matrix would be reduced from 53 to 41, i.e., $\hat{\mathbf{X}}_L \in R^{21 \times 41 \times 300}$. Then steps 2-9 of the procedure described in section 2.2 were implemented on the matrix $\hat{\mathbf{X}}_L$. Figure 9 shows how this feature reduction can substantially decrease the computational effort of the *3*NN imputation. In addition, Figure 10 demonstrates the effect of this feature selection on the accuracy of the different fault diagnosis approaches examined. Generally, higher accuracy and lower CPU time are achieved by feature selection regardless of imputation methods (Appendix A). Therefore, significant improvements can be achieved by appropriate feature selection, although further investigation is required to assess and compare other feature selection techniques.

4.3. Comparison assessment of TEP

In sections 4.1 and 4.2, different algorithms were evaluated using various criteria in the presence of missing measurements. This section presents an overall comparison of the methods through the following normalized indexes:

$$A = \frac{Accuracy - Accuracy^{\min}}{Accuracy^{\max} - Accuracy^{\min}} \tag{17}$$

$$T_i = 1 - \frac{CPU_{imputation} - CPU_{imputation}^{\min}}{CPU_{imputation}^{\max} - CPU_{imputation}^{\min}} \quad (18)$$

$$T_c = 1 - \frac{CPU_{classification} - CPU_{classification}^{\min}}{CPU_{classification}^{\max} - CPU_{classification}^{\min}} \quad (19)$$

$$S = 1 - \frac{Sensitivity - Sensitivity^{\min}}{Sensitivity^{\max} - Sensitivity^{\min}} \quad (20)$$

$$R = \frac{Robustness - Robustness^{\min}}{Robustness^{\max} - Robustness^{\min}} \quad (21)$$

In each index, minimum and maximum performance criteria correspond to the best and worst results calculated in previous sections. In this way, the capability of each tool in terms of each index is normalized and scaled to facilitate the comparison. The spider plots of indexes for each integrated methodology based on the significant feature matrix are illustrated in Figure 11, in which rows and columns correspond to the classification and imputation methods, respectively. Each axis of the spider plots represents an index (Eqs.17-21). For comparative assessment, it should be noted that closeness of indexes to one or zero reflects superiority and inferiority of a correspond method, respectively.

The selection of an appropriate fault diagnosis methodology depends on the process monitoring requirements, but accuracy is imperative because a false alarm may mislead operators. Hence, centroid and BN are not recommended in terms of accuracy. On the other hand, C4.5-*3*NN guarantees the highest accuracy (Eq. 17), and is applicable while FD computational time is not important issue, e.g. in case of offline monitoring, root case analysis or availability of a high-performance computing system. Otherwise, a flexible FD approach enabling the operator to select and switch imputation methods could be considered. In this way, the classification combined with PCA imputation allows quickly inferring the state of the system (Eq. 18), and more reliable results would be obtained through *3*NN imputation. Thus, the GNB, which is too sensitive to imputation values, is not an appropriate choice (Eq. 20). Furthermore, SVM performs better than C4.5 and GNB in terms of robustness (Eq. 21), which is an important issue in case of increasing loss of data. Finally, Figure 12 demonstrates an FD scheme that can tolerate missing data for a general application. It represents interactions between various blocks including: feature extraction, classification and imputation.

5. INDUSTRIAL CASE STUDY

Data from an industrial gas sweetening unit was used for further validation and discussion of the proposed model. Most gas processing plants include a sweetening unit for removing sour gas components from the gas stream, using chemical solvents such as amines (Fig. 13). The acid gas constituents (H_2S and CO_2) react with an aqueous solution in a high-pressure absorber. Subsequently, the acid constituents are stripped from the solvent in a regenerator at high temperature.

One of the most frequent problems in a gas sweetening unit is amine foaming in the absorber, which results in loss of proper vapour-liquid contact, solution hold up and poor solution distribution. The adverse consequences include off-specification product, excessive amine loss, reduced gas-treating capacity and energy loss. Some root causes of foaming, which are considered as faults, are accumulation of heavy hydrocarbon, solid particle in amine and surface active agents in the feed fluid. Therefore, in this work there are four different states, including the mentioned faults and the normal case.

Records of 48 on-line sensors in significant parts of a gas sweetening unit were available from a gas refinery. Among them, 3, 5, 8 and 11 sensors for pressure, level, flow and temperature were selected based on $\mathbf{CI}$ (Eq. 13). The database, $\mathbf{X_L} \in R^{4\times27\times1250}$, was provided by the 27 sensors and consisted of 1250 sample points for each individual state of the system, which include normal, presence of surface active agents, solid particles in the system, and hydrocarbon accumulation in the column. The time interval between samples was 1 min.

A training dataset, $\mathbf{X_L} \in R^{4\times27\times625}$, which had complete records of measurements, was selected. In order to evaluate the proposed procedure, five different testing subsets, $\mathbf{X} \in R^{4\times27\times125}$, were considered. Two test datasets suffered 11% and 26% missing data during operation of the real plant. It is worth mentioning that missing data were not artificially induced. Moreover, artificial incomplete data is required to provide a reference for evaluation of the real mechanism. Thus, 0%, 11% and 26% measurements were randomly deleted from the other three complete test datasets.

6. RESULTS AND DISCUSSION OF THE INDUSTRIAL CASE STUDY

In the TEP case study, missing data had a random mechanism. Herein, it was intended to evaluate the diagnostic performance when dealing with a real case. The accuracy of prediction of missing data and FD performance extremely depend on the informative level of data. When some features highly depend on each other or their duplication or partial redundancy exist, the imputation methods may accurately estimate missing values based on the available values (Kadlec et al., 2009). Consequently, discrimination efficiency of the FD system is not expected to significantly degrade if redundant features exist. In order to assess this important issue, an index, called redundancy ratio, is introduced here.

In order to calculate the redundancy level of a dataset, it is required to evaluate dependency of variables. It is usually characterized in terms of mutual information (Sayood, 2012), which is obtained based on the training dataset as follows:

$$MI_{ij} = \sum_{i=1}^{V}\sum_{j=1}^{V}\sum_{f=1}^{F}\sum_{t=1}^{\tilde{\gamma}} P(x_{it}^{f}, x_{jt}^{f}) \log\left[\frac{P(x_{it}^{f} \mid x_{jt}^{f})}{P(x_{it}^{f})}\right] \qquad MI \in R^{V\times V} \tag{22}$$

Thereafter, inherent informative level of an original database in terms of redundancy can be quantified by the redundancy index, *RI*, which has been introduced by Peng et al. (2005):

$$RI = \frac{1}{|\mathbf{X}|^2} \sum_{x_{it}, x_{jt} \in \mathbf{X}} MI_{ij} \tag{23}$$

In case of incomplete dataset, the redundant information is disturbed. Degradation of informative level of data due to missing data can be characterized by the redundancy ratio, *RR*, as follows:

$$RR = \frac{|\mathbf{X}|^2 \sum_{x_{it}, x_{jt} \in \mathbf{X}_{NaN}^{c}} MI_{ij}}{|\mathbf{X}_{NaN}^{c}|^2 \sum_{x_{it}, x_{jt} \in \mathbf{X}} MI_{ij}} \qquad \mathbf{X}_{NaN}^{c} \cup \mathbf{X}_{NaN} = \mathbf{X} \tag{24}$$

where $\mathbf{X}_{NaN}$ and $\mathbf{X}_{NaN}^{c}$ are incomplete and complete subsets of $\mathbf{X}$.

This analysis was implemented on the dataset of the sweetening unit. $MI \in R^{27\times27}$ was determined based on the training subset using Eq. (22). Then, redundancy ratios of the complete and incomplete testing subsets were achieved based on Eq. (24) which are shown in Figure 14 by solid lines. The negative slopes of these lines show that the missing data degrades the informative level of the data (have measured in terms of redundancy ratio). Furthermore, the effect of the real mechanism of missing data was more adverse than that of the random mechanism which led to the lowest redundancy ratio. It is expected that estimation of missing data might be burden with lower informative level of the real case study. Thereafter, incomplete data was edited by the *3*NN imputation, which is an efficient approach according to the TEP case study (Fig. 11). Then, accuracy of the *3*NN imputation in terms of PAC was assessed, as shown in Figure 14 by dashed lines. Although the real mechanism of missing data has led to a lower PAC in comparison with the random mechanism, the impact is minor.

Next, the FD system was developed based on $\dot{\mathbf{X}}\mathbf{L}$ using C4.5-*3*NN, which was shown to be the most accurate classifier among the others (refer to Section 4.3). Performance of the FD system for real and random mechanisms of missing data was evaluated in terms of accuracy and the results are presented in Figure 15. The minor deviation of accuracy due to mechanisms reveals that this industrial case study has approximately a random mechanism. Consequently, the cause of missing data is ignorable. Therefore, the results of TEP case study can be generalized to this case as well.

7. CONCLUSION

Fault diagnosis of chemical process systems with missing data, which is a common problem in the industrial practice, was investigated. Machine learning provides tools to cope with this challenge rather than ignoring incomplete observations, but they need to be assessed and compared. This work undertakes this comparative study using the TEP benchmark, for which missing values were artificially produced in the datasets.

First, the BN, as a promising exploiting option, was evaluated. This technique can directly infer the state of the system even in case of incomplete information. However, the alternatives of editing the incomplete observations using imputation were shown to produce better performance. Then, the combination of various classifiers -centroid, SVM,

GNB, BN, and C4.5- with imputation methods –mean, PCA, ANN and *k*NN- was investigated. It was found that C4.5 integrated with *3*NN imputation results in the highest accuracy. BN and centroid are not appropriate selection in terms of accuracy. The combination of SVM with *3*NN is highly robust to missing measurements. In addition, SVM was shown to be scarcely sensitive to the imputation method, while the GNB is very sensitive.

The trade-off between performance indicators, including accuracy, robustness, sensitivity and computational effort, was discussed and practical guidelines are proposed. However, the best approach should be selected based on the most important requirements of each practical application. Finally, the feature reduction, by means of the proposed contribution index ($\mathbf{CI}$), was examined to reduce the computation effort, and the promising results obtained encourage future work to assess and compare further feature selection techniques.

Sweetening gas unit of an industrial plant was studied to explore the effect of mechanism of missing data. The real incomplete datasets have a low deviation from the random missing mechanism in terms of redundancy ratio. Thus, the original cause of missing data can be ignored in the analysis. Therefore, results and guidelines obtained in the TEP case study can be also implemented for this case.

NOMENCLATURE

A	normalized index of accuracy
CI_v^f	contribution index
f	index of states of system
F	number of states of system
I	set of variables
J	set of faults
N	number of missing values
R	normalized index of robustness
S	normalized index of sensitivity
$\tilde{S}$	maximum number of significant features
T^f	number of samples
$\tilde{T}$	number of training samples

T_c	normalized index of CPU time of classification
T_i	normalized index of CPU time of imputation
v	index of variables
V	number of variables
$\mathbf{V}^f$	set of significant features for each fault
$\mathbf{V}$	set of significant features
x	measured variable
$\dot{.}$	standardized value of measurement
$\dot{.}$	imputed values of missing variable
$\tilde{\mathbf{X}}_L$	labelled process dataset
$\mathbf{X}_L$	training matrix
$\dot{\mathbf{X}}_L$	standardised training matrix
$\mathbf{X}$	test matrix
$\dot{\mathbf{X}}$	standardised test matrix
$\dot{\mathbf{X}}$	estimated test dataset
$\hat{\mathbf{X}}_L$	significant feature matrix

Greek symbols

μ_v	mean of the v^{th} variable in the training dataset
$\dot{\mu}$	mean of true values of the v^{th} variable in the test dataset
$\dot{\mu}$	mean of a imputed the v^{th} variable in the test dataset
σ_v	variance of the v^{th} variable in the training dataset

Acronyms

ANN	artificial neural network
BN	Bayesian network
Cond	condensate
CPU	central processing unit
CWS	cold water stream
DCS	distributed control system
EM	expectation–maximization
FD	fault diagnosis

FI	flow indicator
FIS	fuzzy inference system
GNB	Gaussian naïve Bayes
ID3	Iterative Dichotomiser 3
*k*NN	*k*-nearest neighbourhood
LI	level indicator
MAR	missing at random
MCAR	missing completely at random
MI	mutual information
ML	maximum-likelihood
MLP	Multi-layer perceptron
NMAR	not missing at random
PAC	predictive accuracy index
PCA	principal component analysis
PI	pressure indicator
PLC	programmable logic controller
PLS	partial least squares
SC	sample connection
SCADA	supervisory control and data acquisition
Stm	steam
SVM	support vector machines
TEP	Tennessee-Eastman process
TI	temperature indicator

ACKNOWLEDGEMENTS

Financial support from the Iranian ministry of science, research and technology, as well as South Pars Gas Complex is acknowledged. Also, Financial support from the Spanish MICINN/MINECO and FEDER funds (Projects EHMAN, DPI2009-09386 and SIGERA, DPI2012-37154-C02-01) and the Generalitat de Catalunya (2014 SGR 1092) is fully appreciated.

REFERENCE

Abdella M, Marwala T. The use of genetic algorithms and neural networks to approximate missing data in database. ICCC 2005: Proceedings of 3rd International Conference on Computational Cybernetics; April 13-16; Mauritius:IEEE; 2005. p. 207-12.

Ahmad S, Tresp V. Some solutions to the missing feature problem in vision. In: Hanson SJ, Cowan JD, Giles CL, editors. Advances in neural information processing systems. San Mateo, CA.: Morgan Kaufmann; 1993. p. 393-401.

Amarnath M, Sugumaran V, Kumar H. Exploiting sound signals for fault diagnosis of bearings using decision tree. Measurement 2013; 46(3):1250-6.

Atkeson CG, Moore AW, Schaal S. Locally weighted learning for control. In: Aha DW, editor. Lazy learning. Netherlands: Springer; 1997. p. 75-113.

Auffarth B, López M, Cerquides J. Comparison of redundancy and relevance measures for feature selection in tissue classification of CT images. ICDM2010: Proceeding of International Conference on Data Mining; Dec. 13-17; Springer; 2010. p. 248-62.

Bastin G, Dochain D. On-line estimation and adaptive control of bioreactors. New York, Amsterdam: Elsevier; 1990.

Batista GE, Monard MC. A study of K-nearest neighbour as an imputation method. In: Abraham A, editor. Hybrid Intell Syst. IOS Press; 2002. p. 251-60.

Bishop CM. Neural networks for pattern recognition. New York: Oxford university press; 1995.

Chen JM, Chen BS. System parameter estimation with input/output noisy data and missing measurements. IEEE Transactions on Signal Processing 2000; 48(6):1548-58.

Chéruy A. Software sensors in bioprocess engineering. J Biotechnol 1997; 52(3):193-9.

Cristianini N, Taylor JS. An introduction to support vector machines and other kernel-based learning methods. UK: Cambridge university press; 2000.

Duda RO, Hart PE, Stork DG. Pattern classification.2nd ed. New York: John Wiley & Sons; 2001.

Fortuna L, Graziani S, Rizzo A, Xibilia MG. Soft sensors for monitoring and control of industrial processes. Italy: Springer Science & Business Media; 2007.

Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Mach Learn 1997; 29(2-3):131-63.

Gabrys B. Neuro-fuzzy approach to processing inputs with missing values in pattern recognition problems. Int J Approximate Reasoning 2002; 30(3):149-79.

García-Laencina PJ, Sancho-Gómez JL, Figueiras-Vidal AR. Pattern classification with missing data: a review. Neural Comput Appl 2010; 19(2):263-82.

Ge Z, Song Z. Robust monitoring and fault reconstruction based on variational inference component analysis. J Process Control 2011; 21(4):462-74.

Gonzalez G. Soft sensors for processing plants. IPMM99: Proceedings of the 2nd International Conference on Intelligent Processing and Manufacturing of Materials; July 10-15; USA: IEEE; 1999. p. 59-69.

Huang B. Bayesian methods for control loop monitoring and diagnosis. J Process Control 2008; 18(9):829-38.

Ibrahim JG. Incomplete data in generalized linear models. J Am Stat Assoc 1990; 85(411):765-9.

ISO. 7498-1. Information Technology–Open Systems Interconnection–Basic reference model; IEC: 1994. p.

Ji C, Elwalid A. Measurement-based network monitoring and inference: scalability and missing information. Selected Areas in Communications 2002; 20(4):714-25.

Jiang J, Marron J, Jiang X. Robust centroid based classification with minimum error rates for high dimension, low sample size data. J Stat Plan Inference 2009; 139(8):2571-80.
Kadlec P, Gabrys B, Strandt S. Data-driven soft sensors in the process industry. Comput Chem Eng 2009; 33(4):795-814.
Kadlec P, Grbić R, Gabrys B. Review of adaptation mechanisms for data-driven soft sensors. Comput Chem Eng 2011; 35(1):1-24.
Little RJ, Rubin DB. Statistical analysis with missing data. New Jersey: John Wiley & Sons; 2014.
Liu H, Buvat JC, Estel L, Polaert I. Bayesian network method for fault diagnosis in a continuous tubular reactor. Chem Produc Process Mod 2010; 5(1):1-11.
Luo JX, Shao HH. Developing soft sensors using hybrid soft computing methodology: a neurofuzzy system based on rough set theory and genetic algorithms. Soft Computing 2006; 10(1):54-60.
MacGregor J, Cinar A. Monitoring, fault diagnosis, fault-tolerant control and optimization: Data driven methods. Comput Chem Eng 2012; 47:111-20.
Marquez L, Escudero G, Martinez D, Rigau G. supervised corpus-based methods for WSD. In: Agirre E, Edmonds PG, editors. Word sense disambiguation: Algorithms and applications. Springer; 2007. p. 167-216.
Nelson PR, MacGregor JF, Taylor PA. The impact of missing measurements on PCA and PLS prediction and monitoring applications. Chemom Intell Lab Syst 2006; 80(1):1-12.
Nelson PR, Taylor PA, MacGregor JF. Missing data methods in PCA and PLS: Score calculations with incomplete observations. Chemom Intell Lab Syst 1996; 35(1):45-65.
Palit AK, Popovic D. Computational Intelligence in Time Series Forecasting. Springer Science & Business Media; 2006. 393 p.
Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann; 1988. 552 p.
Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. Pattern Analysis Mach Intell 2005; 27(8):1226-38.
Quinlan JR. C4. 5: programs for machine learning. Morgan kaufmann; 1993. 302 p.
Rodriguez A, Gatrell J, Karas J, Peschke R. TCP/IP Tutorial and Technical overview. Citeseer; 2002. 975 p.
Rubin DB. Inference and missing data. Biometrika 1976; 63(3):581-92.
Salton G. Automatic text processing: the transformation, analysis, and retrieval of information by computer. Addison-Wesley; 1989. 530 p.
Sayood K. Introduction to data compression. Newnes; 2012. 740 p.
Schafer JL. Analysis of incomplete multivariate data. CRC press; 1997. 448 p.
Schapire RE, Singer Y. Improved boosting algorithms using confidence-rated predictions. Mach Learn 1999; 37(3):297-336.
Scheffer J. Dealing with missing data. Res Lett Inf Math Sci 2002; 3:153-60.
Sharpe PK, Solly R. Dealing with missing values in neural network-based diagnostic systems. Neural Comput Appl 1995; 3(2):73-7.
Venkatasubramanian V, Rengaswamy R, Kavuri SN, Yin K. A review of process fault detection and diagnosis: Part III: Process history based methods. Comput Chem Eng 2003; 27(3):327-46.

Verron S, Tiplica T, Kobi A. Multivariate control charts with a Bayesian network. ICINCO 2007: Proceeding of 4th International Conference on Informatics in Control, Automation and Robotics; May 9-12; France; 2007.
Walczak B, Massart D. Dealing with missing data: Part I. Chemom Intell Lab Syst 2001a; 58(1):15-27.
Walczak B, Massart DL. Dealing with missing data: Part II. Chemom Intell Lab Syst 2001b; 58(1):29-42.
Wang S. Classification with incomplete survey data: a Hopfield neural network approach. Comput Oper Res 2005; 32(10):2583-94.
Wang Y, Yang Y, Zhou D, Gao F. Active fault-tolerant control of nonlinear batch processes with sensor faults. Ind Eng Chem Res 2007; 46(26):9158-69.
Warne K, Prasad G, Rezvani S, Maguire L. Statistical and computational intelligence techniques for inferential model development: a comparative evaluation and a novel proposition for fusion. Eng Appl Artf Intell 2004; 17(8):871-85.
Yélamos I, Escudero G, Graells M, Puigjaner L. Performance assessment of a novel fault diagnosis system based on support vector machines. Comput Chem Eng 2009; 33(1):244-55.
Yélamos I, Graells M, Puigjaner L, Escudero G. Simultaneous fault diagnosis in chemical plants using a multilabel approach. AIChE J 2007; 53(11):2871-84.
Yin S, Ding SX, Haghani A, Hao H, Zhang P. A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process. J Process Control 2012; 22:1567– 81.
Yoon SY, Lee SY. Training algorithm with incomplete data for feed-forward neural networks. Neural Proc Lett 1999; 10(3):171-9.
Zhao C, Fu Y. Statistical analysis based online sensor failure detection for continuous glucose monitoring in type I diabetes. Chemom Intell Lab Syst 2015; 144:128-37.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version.

List of Tables:

List of Figures: