Expert knowledge-guided feature selection for data-based industrial process monitoring.

Selección de variables guiada por conocimiento del experto para el monitoreo basados en datos de procesos industriales

Cesar Uribe, Claudia Isaza[*]

Department of Electronic Engineering. Universidad de Antioquia. Calle 67 N°. 53-108 Bl.19 Of. 426, Medellín, Colombia.

(Recibido el 10 enero de 2012. Aceptado el 6 noviembre del 2012)

Abstract

Industrial processes are characterized to be in open environments with uncertainty, unpredictability and nonlinear behavior. Rigorous measuring and monitoring is required to strive for product quality, safety and finance. Therefore, data-based monitoring systems have gain interest in academia and industry (e.g. clustering). However industrial processes have high volumes of complex and high dimensional data available, with poorly defined domains and sometimes redundant, noisy or inaccurate measures with unknown parameters. When a mechanistic or structural model is not available or suitable, selecting relevant and informative variables (reducing the high dimensionality) eases pattern recognition to identify functional states of the process. In this paper, we address the feature selection problem in data-based industrial processes monitoring where a mathematical or structural model is not available or suitable. Expert knowledge-guidance is used inside a wrapper feature selection based on clustering. The reduced set of features is capable of represent intrinsic historical-data structure integrating the expert knowledge about the process. A monitoring system is proposed and tested on an intensification reactor, the 'open plate reactor (OPR)', over the thiosulfate and the esterification reaction. Results show fewer variables are needed to correctly identify the process functional states.

---------- *Keywords*: Feature selection, processes monitoring, fault detection, fuzzy clustering

Resumen

Los procesos industriales se caracterizan por estar en ambientes abiertos, inciertos y no lineales. La medición y monitoreo de estos busca calidad,

[*] Autor de correspondencia: teléfono: 57 + 4 + 219 85 60, fax: 57 + 4 + 219 55 84, correo electrónico: cisaza@udea.edu.co (C. Isaza)

seguridad y economía en los productos. Los sistemas de monitoreo basados en datos han ganado un gran interés en la academia y en la industria, pero los procesos industriales tienen grandes volúmenes de datos complejos y de alta dimensión, con dominios poco definidos, medidas redundantes, ruidosas e imprecisas y parámetros desconocidos. Cuando un modelo mecánico no está disponible, seleccionar las variables relevantes e informativas (reduciendo la dimensión de los datos) facilita la identificación de los patrones en los estados funcionales del proceso. En este artículo se propone usar el conocimiento del experto como guía dentro de un *wrapper* de selección de descriptores basado en agrupamiento para reducir el conjunto de variables necesarias para representar la estructura intrínseca de los datos históricos del proceso. Un sistema de monitoreo es propuesto y evaluado en un reactor de intensificación, el *Open Plate Reactor,* en las reacciones de tiosulfato y esterificación. Los resultados muestran que sólo algunas variables son necesarias para identificar correctamente los estados funcionales del proceso.

---------- *Palabras clave*: Selección de variables, monitoreo de procesos, detección de fallos, agrupamiento difuso.

Introduction

Large volumes of complex and high dimensional data available set a barrier for developing efficient decision support and monitoring systems [1]. Using relevant and informative variables eases data understanding, classification accuracy and computational efficiency [2], [3]. For example, Mukse et al. [4] used the Pareto optimal trade-off between the process information that can be obtained and the sensor cost for the selected process measurements, but a process model is needed. Sikora et al. [5] designed an effective and efficient genetic algorithm for a wrapper feature selection method based on Hausdorff distance measure in a supervised manner. Fraleigth et al. [6] developed a sensor system selection for model-based real-time optimization. Verron et al. [7] proposed supervised fault diagnosis with feature selection based on discriminant analysis and mutual information. Bensch et al. [8] tackled the problem of identifying the features responsible for success or failure in the manufacturing process in a supervised context. These methods focus on constructing process models and identify the gap with the actual system using supervised learning. However, complex processes do not always have classical models available [9]. Thus, several researchers focused on the development of robust and reliable monitoring systems based on data analysis.

Data-based monitoring systems use measurement's information to identify process behaviors as functional states or classes. Such information is classified according to its resemblance with previously classified historical data [10]. However, in industrial processes, class labels are unknown and most of the knowledge is held by the expert. Such knowledge constrains knowledge discovery, avoid the data over fitting problem [11] and describes the relationship between attributes, categories and correlations among them. The expert judgment approach may result in an effective feature selection without bias by the distribution of the training set [12]. Real-life applications require the involvement of domain experts to validate the allocation of operating states of the process into classes resulting from clustering. Nevertheless, high dependency upon expert knowledge is not desirable due to their inability to examine large amounts of data in a rigorous fashion without the effects of boredom or frustration [13]. Using computational intelligence techniques seems to be an alternative to take into account the process expert knowledge. In this context, techniques that use data artificially labed by the expert are valuable to diagnosis and classification systems. [14].

In this paper, a wrapper feature selection guided by the process expert's knowledge is proposed. Expert's knowledge is not used for supervised training but as guidance in order to look for clustering results as similar as the expert data partition maintaining a cluster structure. The method is applied on fault detection and monitoring (i.e. classification of the process dynamic in a predefined functional state) of the ‘open plate reactor (OPR)’ [15, 16] on the thiosulfate reaction and the esterification reaction.

Next section shows the proposed wrapper framework for feature selection: feature search, clustering algorithms, clustering quality assessment. Third section details the open plate reactor application over two chemical reactions (esterification and thiosulfate). Results are presented in section four. Last section shows conclusions and future work.

Wrapper feature selection guided by the expert knowledge

The wrapper methodology [2], offers a simple and powerful way to address the problem of variable selection [17], regardless of the chosen learning machine or quality subset criterion [18]. The performance of the induction algorithm guides the search, producing better results than filter feature selection methods for specific applications [19].

Figure 1 shows a detailed graphic of the proposed methodology. Historical data (i.e. database of the process) is defined in the $N \times n$ space, as a set Ω; N is the number of elements, n is the number of features in the original feature set $F \in \Re^n$ and $F_r \in \Re^r$ with $r \leq n$ represented as $\Omega \downarrow F_r$. The clustering algorithm partitions the data subset into c clusters, optimizing some metric J over the data. Consider the clustering algorithm as $Y = J(\Omega \downarrow F_r, \lambda)$ where λ are the clustering method parameters. Let $Y_\omega^T = [y_1, y_2, \dots, y_N]$, $y_i \in \{1, 2, \dots, c\}$ be the partition produced by the clustering algorithm and $Q(Y_\omega, Y_\varepsilon) = \varphi$ be the performance function that assesses similarity between two partitions (e.g. expert and clustering partition). The feature search procedure generates the optimal set of features F_{OP} by testing different forms of the map $\Omega_r = f(\Omega, \varphi)$.

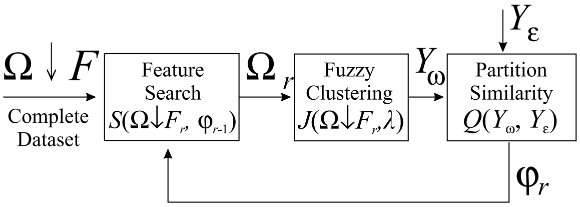

Figure 1 Proposed wrapper method based on clustering

feature search

Finding the optimal feature subset F_{OP} requires either an exhaustive search that involves the evaluation of 2^n subsets (becoming infeasible since n is large) [19] or the monotonicity of a pertinence measure. Two different sequential search strategies were implemented to analyze the case study: Sequential Forward Selection (SFS) and Sequential Backward Elimination (SBE). SFS starts with the subset $F_{0,,}$ n partitions are obtained using clustering and its quality is computed. First, each feature subset includes only one variable. The feature subset Ω_1 associated with the highest quality φ, is set to be the first selected variable in the vector v. Each feature that is not yet included in v is included and the quality of the n - 1 partitions is computed. The vector v with two features that led to the highest quality is selected as the new vector of selected features. These steps are repeated, adding one feature per iteration until a pre-specified number of characteristics is achieved (e.g. the total number of characteristics) or a performance criterion is met. Sequential Backward Elimination makes the search in the opposite direction. Starting with the full set of features, at each step the features are removed one by one.

Clustering algorithm

Data-based monitoring systems based on clustering try to find similarities in the process data and group them into classes that correspond to functional states. The term “similarity” should be understood as a mathematical measure of similarity, in some well-defined sense (e.g. distance based, hierarchy based, possibility based among others). In crisp clustering, when a data partition is build, a single sample belongs to only one cluster. The fuzzy clustering extends this notion, and each data belongs to all clusters with different membership degrees.

In this article the Learning Algorithm for Multivariate Data Analysis (LAMDA) is used. LAMDA method has been widely used in the literature for the construction of systems for monitoring industrial processes [14, 16, 17, 20-24]. LAMDA [25] is based on finding the overall adequacy level of each individual to each class, called Global Adequacy Degree (GAD). The GAD is the membership degree of each object to each class. Its value is estimated using the contributions of the features based on a marginal concept of adequacy which replaces the use of traditional distance approximations. The contribution of each descriptor is called the Marginal Adequacy Degree (MAD) and it is computed using a possibility function. The class adequacy concept is expressed as the "fuzzy" truth value of a compound sentence using logical connectives between elementary assertions. Attributes can be numeric, symbolic or mixed (which is an advantage compared to other fuzzy classifiers that can only handle numeric descriptors). Also, LAMDA methodology does not require a number of classes to be specified as parameter, thus, it is capable of producing a data partition estimating the number of classes based on the data distribution. For a complete description of the LAMDA methodology see [25, 26].

Feature evaluation criteria

Partitions results are evaluated comparing the clustering algorithm and the process expert partition. The expert's partition is not used as classification vector in supervised way because even though the proposed method looks for producing partitions similar to the expert proposal, it still looks for finding underlying structures among data in order to identify similarities in the historical data [27].

The Index of Dissimilarity *Idn* proposed by Lopez de Mantaras in [28, 29] allows to compare two data partitions with different number of classes and it has been recently used to compare partitions of industrial process [14]. The contingency matrix is established for two partitions: A (whose classes are denoted (a_1, a_2, … , a_i,…, a_p)) and B (whose classes are denoted (b_1 , b_2,…, b_j,…, b_r)). The probabilities corresponding to each class and the probability of the intersection between a class of A partition and a partition class B are noted as Eq.1:

$$P_i = P(a_i),\ P_j = P(b_i),\ P_{ij} = P(a_i \cap b_i) \qquad (1)$$

where $a_i \cap b_i$ is formed by the elements that belong simultaneously to the latter class a_i and class b_i. The probabilities satisfy Eq. 2:

$$P_i = \sum_{j=1}^{r} P_{ij}\,, P_j = \sum_{i=1}^{p} P_{ij}\,, \sum_{j=1}^{r}\sum_{i=1}^{p} P_{ij} = 1 \qquad (2)$$

The probability of elements belonging to this class a_i and class b_i is computed with Eq. 3. *M* is the cardinality *N* and the total number of individuals ordered *M(X*).

$$P(a_i \cap b_i) = \frac{M(a_i \cap b_i)}{N} \qquad (3)$$

The *Idn* is zero only if the contingency matrix is "almost diagonal" or "quasi-diagonalizable", that is, when the partitions are either equal or compatible or equal modulo zero. The *Idn* is estimated from the conditional information between partitions A and B.

A normalized index of dissimilarity *Idn* = *φ* between the clustering partition Y_ω and expert partition Y_ε is defined in Eq. 4.

$$Idn(A,B) = \frac{\bar{H}\left(\frac{Y_\omega}{Y_\varepsilon}\right)}{\bar{H}\left(\frac{Y_\omega}{Y_\varepsilon}\right)+\bar{H}(Y_\omega,Y_\varepsilon)} \qquad (4)$$

$$= \frac{-\sum_{j=1}^{r}\sum_{i=1}^{p} P_{ij}\, log_2\left(\frac{P_{ij}}{P_i}\right)}{-\sum_{j=1}^{r}\sum_{i=1}^{p} P_{ij} + \sum_{j=1}^{r}\sum_{i=1}^{p} P_{ij}\, log_2(P_{ij})}$$

If the partition Y_ω is consistent or equal to Y_ε, *Idn* = 0 and *Idn* = 1 in the opposite case.

Cases studies: Open Plate Reactor –OPR

The OPR is a plate heat exchanger of new design [15]. One side is used as a chemical continuous reactor while the other side a cooling/heating thermal fluid flows. The primary reactant R_1 flows from the inlet to the outlet of the reactor (see figure 2). The secondary reactant R_2 can then be injected along the reactor side with R_2. Depending on the reaction, the utility flow is used to cool (exothermic reaction) or heat (endothermic reaction) the reactor side.

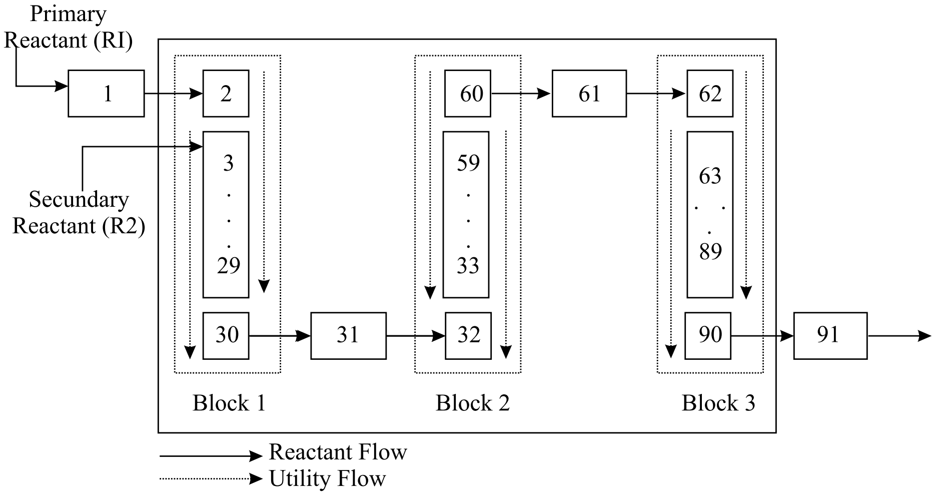

Figure 2 Schematic Representation of the OPR

Figure 2 shows the schematic representation of the pilot plant; two feeding loops ensure the introduction of the reactants in the reactor at normal temperature [15]. The OPR has 27 available sensor measurements from temperatures and pressures from different cells of the reactor.

The OPR is studied under two chemical reactions; thiosulfate and esterification; described below. Failures in the OPR for the thiosulphate reaction and the esterification reaction were introduced in the process in the form of disturbances on the main variables: increase and decrease of temperatures and flows of the utility, primary and secondary reactants and increase and decrease of the compositions of the primary and secondary reactants.

Thiosulphate reaction

The thiosulphate reaction has the following characteristics: its stoichiometry and kinetic are known, the reaction is irreversible, fast and highly exothermic.

Table 1 shows a description of all functional states over the thiosulfate reaction. The database used is composed by the measure of the 27 variables with 17 simulated faults over 2076 time samples. The reaction scheme is in Eq. 5:

$$2Na_2S_2O_3 + 4H_2O_2 \\ \rightarrow Na_2S_3O_6 + Na_2SO_4 + 4H_20 \tag{5}$$

Table 1 OPR, thiosulfate and esterification reaction faults description

Functional State Description				*Thiosulfate*		*Esterification*	
Description	***Fluid***	***Variable***	***Id***	***Initial***	***Final***	***Initial***	***Final***
Normal	-	-	1	-	-	-	-
↓$F(U_f)$	Utility	Flow	2	0.916m^3/h	0.22m^3/h	3m^3/h	1m^3/h
↑$F(U_f)$	Utility	Flow	3	0.916m^3/h	1.76m^3/h	3m^3/h	5m^3/h
↓$T(U_f)$	Utility	Temp.	4	13.37°C	8°C	70°C	60°C
↑$T(U_f)$	Utility	Temp.	5	13.37°C	20°C	70°C	80°C
↓$F(R_1)$	Prim. Reac.	Flow	6	39.38m^3/h	30m^3/h	15m^3/h	10m^3/h
↑$F(R_1)$	Prim. Reac.	Flow	7	39.38m^3/h	50m^3/h	15m^3/h	20m^3/h
↓$T(R_1)$	Prim. Reac.	Temp.	8	19.8°C	10°C	20°C	10°C
↑$T(R_1)$	Prim. Reac.	Temp.	9	19.8 °C	30°C	20°C	30°C
↓$F(R_2)$	Sec. Reac.	Flow	10	9.87m^3/h	5m^3/h	10m^3/h	7m^3/h
↑$F(R_2)$	Sec. Reac.	Flow	11	9.87m^3/h	15m^3/h	10m^3/h	13m^3/h
↓$T(R_2)$	Sec. Reac.	Temp.	12	19.15°C	10°C	20°C	10°C
↑$T(R_2)$	Sec. Reac.	Temp.	13	19.15°C	30°C	20°C	30°C
↑$C(R_1)$	Prim. Reac.	Mol. Frac.	14	0.0157 %	0.0173 %	0.966 %	0.95 %
↓$C(R_1)$	Prim. Reac.	Mol. Frac.	15	0.0157 %	0.0141 %	0.966 %	0.981 %
↑$C(R_2)$	Sec. Reac.	Mol. Frac.	16	0.1289 %	0.15 %	0.994 %	0.99 %
↓$C(R_2)$	Sec. Reac.	Mol. Frac.	17	0.1289 %	0.1 %	0.994 %	0.998 %
↓$sd(R_f)$	Util. Flow	Shutdown	18	0.916m^3/h	0.01m^3/h	-	-

In order to validate the generated model using just the selected subset of sensors (the selected features), a test database with 735 new samples described only by the selected features was simulated. Six new faults were induced in the test dataset as described in table 2.

Table 2 Faults description in the test dataset (thiosulfate)

Fault	*Description*	*Start (t)*	*End (t)*
Fault 1	↓$F(U_f)$: 0.916m^3/h → 0.3m^3/h	15	75
Fault 2	↑$C(R_1)$: 0.0137% → 0.017%	135	195
Fault 3	↑$T(U_f)$: 13.37°C → 15°C	255	315
Fault 4	↑$C(R_1)$: 0.0137% → 0.017%	375	435
	↓$F(U_f)$: 0.916m^3/h → 0.3m^3/h	375	435

Fault	*Description*	*Start (t)*	*End (t)*
Fault 5	↑$C(R_1)$: 0.0137% → 0.0165%	495	555
	↑$F(U_f)$: 0.916m^3/h → 1.5m^3/h	495	555
Fault 6	↑$C(R_1)$: 0.0137% → 0.0165%	615	675
	↓$F(U_f)$: 0.916m^3/h → 0.6m^3/h	615	675
	↑$T(U_f)$: 13.37 → 14.8°C	615	675

Estertification reaction

The esterification reaction is slow and weakly exothermic. To accelerate it, it is necessary to heat the reaction medium. In this case, the utility flow serves as fluid heating. In total, 16 faults have been applied to the reactor. Failures in the OPR are disturbances on the temperatures and flow rates of main reactant0 (C_4H_8O) secondary or injected reactant ($C_6H_{10}O_3$), cooling system (utility), and composition in primary and secondary reagents, see table 2.

Validation on the esterification reaction results is made over a test database consisting of 410 new samples described only by the selected feature was simulated. Five new faults were induced in the test dataset as described in table 3.

Table 3 Faults description in the test dataset (esterification reaction)

Fault	*Description*	*Start (t)*	*End (t)*
Fault 1	↓$C(R_2)$: 0.994% → 0.991%	10	50
Fault 2	↓$T(U_f)$: 70°C → 65°C	90	130
Fault 3	↓$C(R_2)$: 0.994% → 0.991%	170	210
	↑$F(R_2)$: 10m^3/h → 13m^3/h	170	210
Fault 4	↓$F(U_f)$: 3m^3/h → 2m^3/h	250	290
Fault 5	↓$F(U_f)$: 3m^3/h → 2m^3/h	330	370
	↑$T(U_f)$: 70°C → 60°C	330	370

Experimental results and discussion

Variables representing input pressures for primary and secondary reactants were eliminated since they are constant. Feature selection is applied to the remaining 25 variables. The data subset associated with the lowest *Idn* value is represented by the set of features that minimize the dissemblance between the partition produced by the clustering algorithm and the partition proposed by the expert knowledge. For the thiosulfate reaction, the feature set $f_{SFS}^{t}(5) = \{1, 22, 7, 8, 24\}$ and $f_{SBE}^{t}(5) = \{24, 8, 7, 22, 1\}$ are selected as the best set of features reaching $Idn = 0.03232$ and $Idn = 0.03171$ respectively, see figure 3. For the esterification reaction, features sets $f_{SFS}^{e}(8) = \{5, 14, 2, 1, 22, 18, 20, 15\}$ and $f_{SBE}^{e}(7) = \{21, 8, 6, 13, 19, 22, 3\}$ with dissemblance index values of $Idn = 0.04048$ and $Idn = 0.04193$, see figure 4.

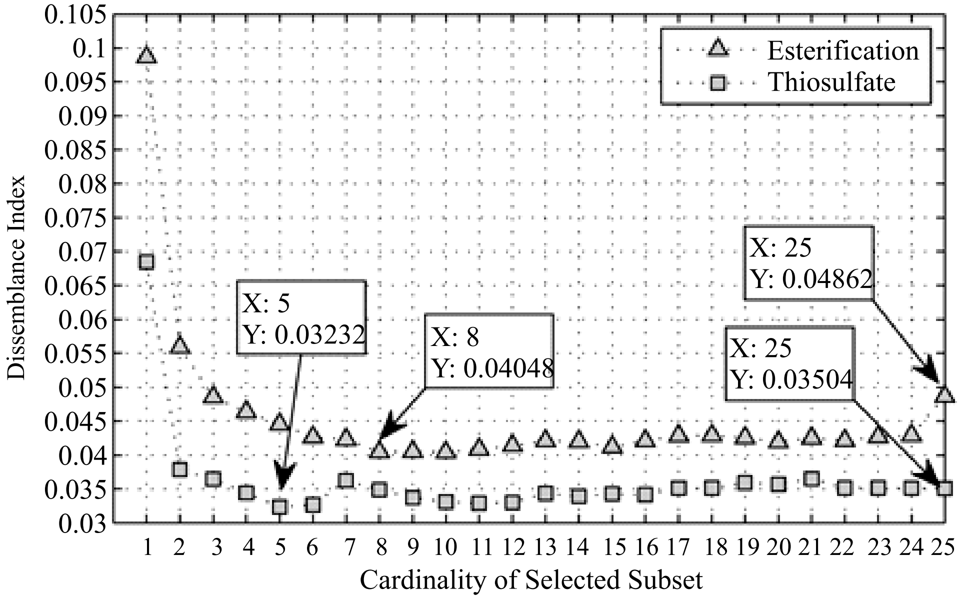

Figure 3 Dissimilarity Index of the OPR reactions studied, with the SFS

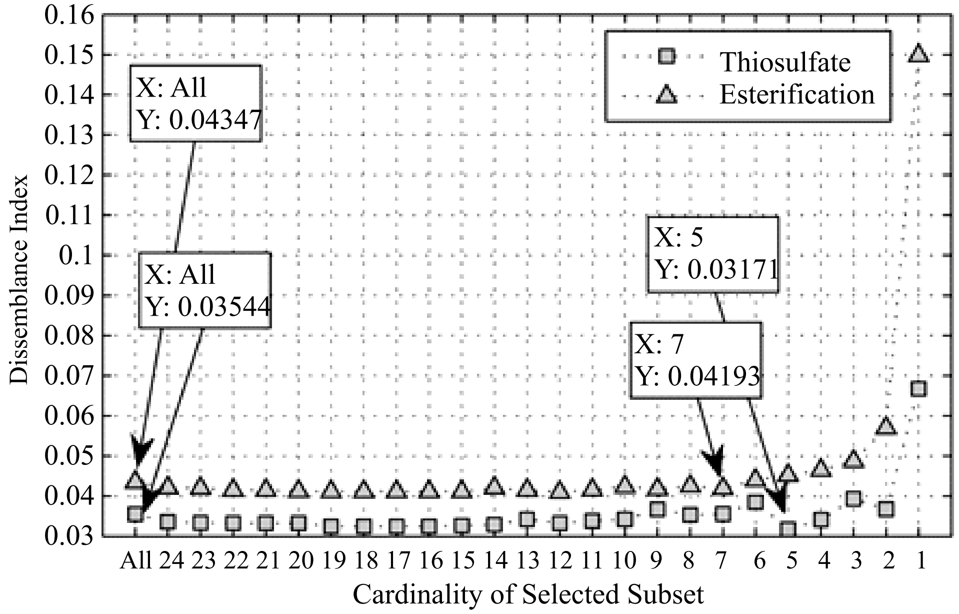

Figure 4 Dissimilarity Index of the OPR reactions studied, with the SBE

Figures 5, 6, 7 and 8 show the classification results of the training datasets when using just the selected features. The monitoring system identifies all functional states for both chemical reactions studied, with similar results for SFS and SBE. Additionally, a new class is defined, the transition class. This class represents a deviation from the Normal state and it is not included by the process expert. False alarms appears at the end of some faults, most of them are misclassification with the increase of Temperature of the Utility Flow $\uparrow T(U_f)$ since the utility flow acts as temperature regulation and influences directly all functional states.

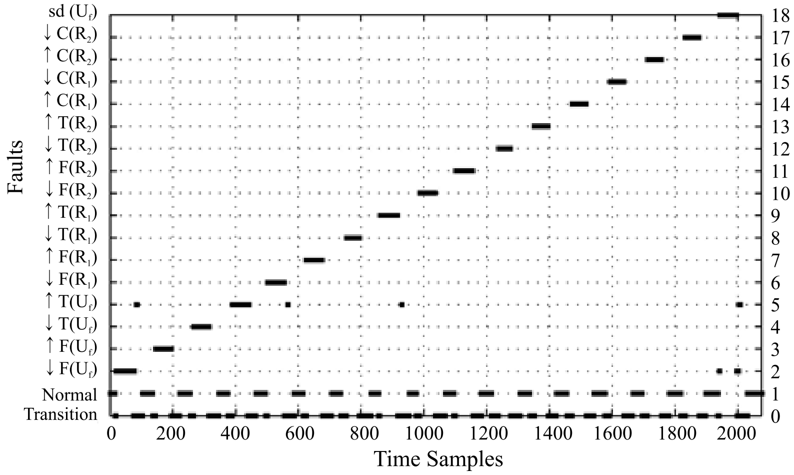

Figure 5 Map of Clustering Results for Thiosulfate Reaction with SFS

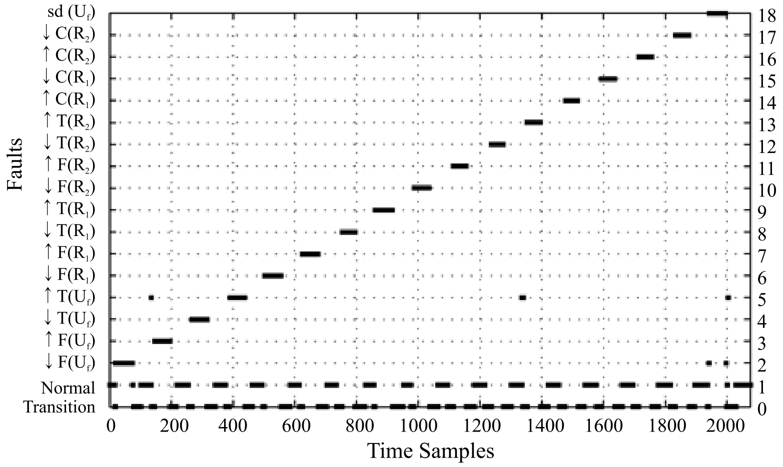

Figure 6 Map of Clustering Results for Thiosulfate Reaction with SBE

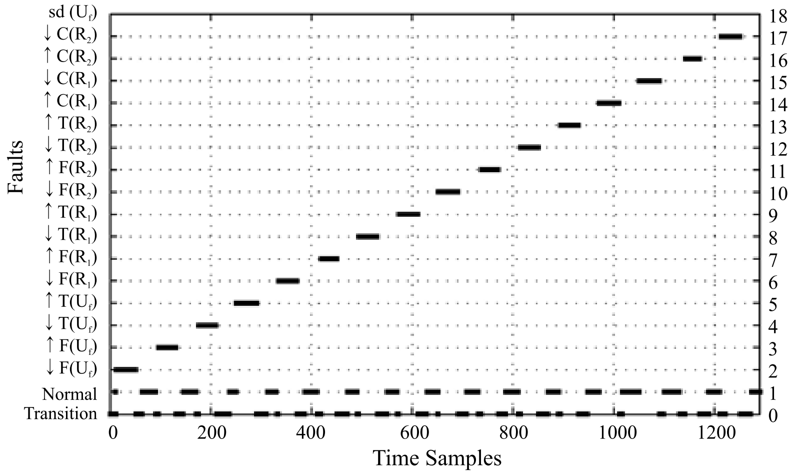

Figure 7 Map of Clustering Results for Esterification Reaction with SFS

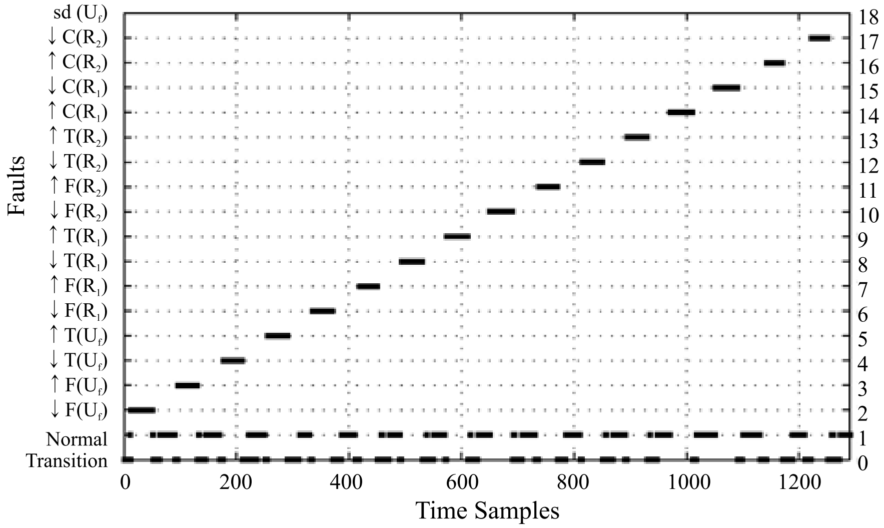

Figure 8 Map of Clustering Results for Esterification Reaction with SBE

The resulting classifiers are tested on validation datasets, obtaining the results shown in figures 9, 10, 11 and 12. For the thiosulfate reaction, when using SFS, the first three single disturbances are correctly identified. The classifier is able to identify the fault when several disturbances are presented simultaneously. Perturbation 5 is classified as normal because the combined effect of both perturbations cancels out. The reactor is fed with more primary reactant, but the utility fluid cools more, which corresponds to a normal operating state. For the esterification reaction both procedures, SFS and SBE, produce different sets of features. Fault 4 is identified as normal in both cases, since the esterification reaction is very exothermic, so the impact of such small variation does not affect la reaction. In the SBE search, the second perturbation corresponding to $\downarrow T(U_f)$, is misclassified with functional state $\uparrow F(R_2)$ this is because a decrease on the utility fluid temperature increases the temperature of the reaction, and this increase appears when there is an increase of flow of the Secondary Reactant.

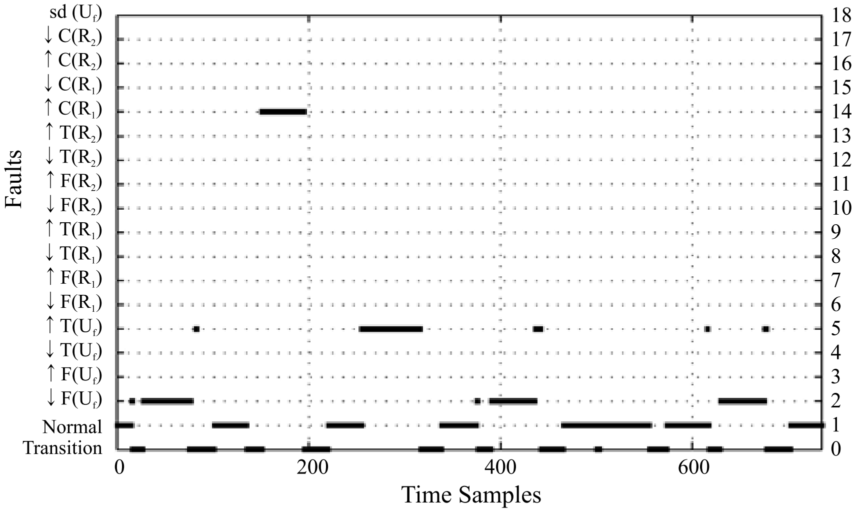

Figure 9 Map of Clustering for the Test Dataset of the Thiosulfate Reaction with SFS

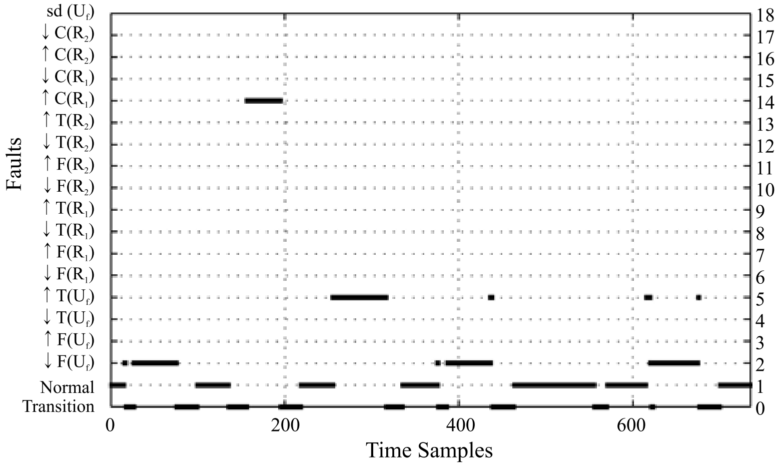

Figure 10 Map of Clustering for the Test Dataset of the Thiosulfate Reaction with SBE

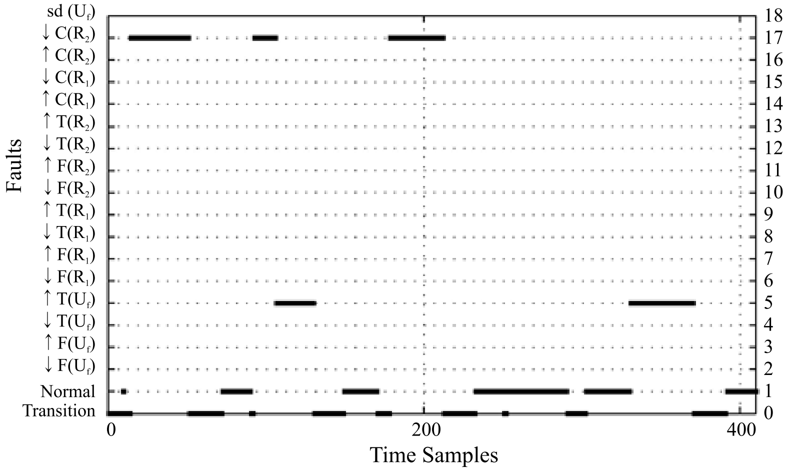

Figure 11 Map of Clustering for the Test Dataset of the Esterification Reaction with SFS

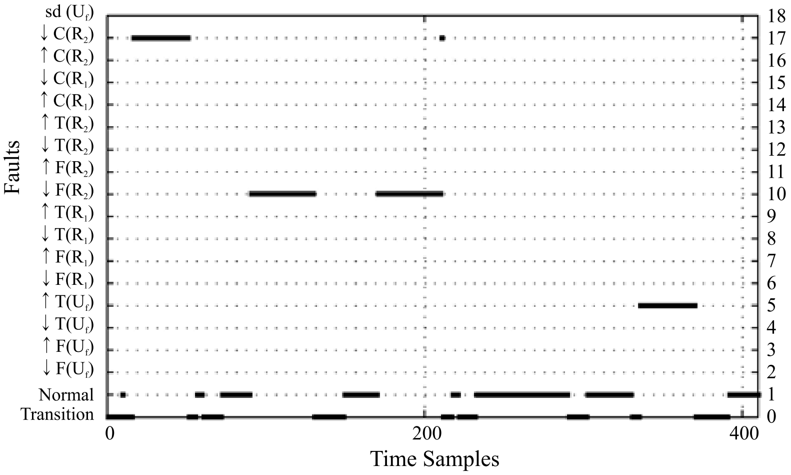

Figure 12 Map of Clustering for the Test Dataset of the Esterification Reaction with SBE

Previously in [16], the authors proposed a ranking method based on information-theoretic measures to evaluate the amount of information within each variable to select the most informative ones. Additionally, [17] and [30] explore wrapper approaches for unsupervised feature selection. Tables 4 and 5 show a comparison of previous feature selection results on the same process showing a better performance, with lower *Idn* value

Table 4 Comparison of dissimilarity with previous proposals for the thiosulfate reaction

Author	***Feature Set***	***Idn***	***Type***
Orantes et al. [16]	[1, 4, 5, 6, 12, 14, 22, 25, 26]	37.49	Filter
Uribe et al. [17]	[1, 7, 25, 9, 6, 21, 20, 8, 10]	39.38	Wrapper
Uribe et al. [30]	[6, 8, 1, 7, 25]	34.89	Wrapper
Expert-Guided SFS	[1, 22, 7, 8, 24]	32.32	Wrapper
Expert-Guided SFS	[24, 8, 7, 22, 1]	31.71	Wrapper

Table 5 Comparison of dissimilarity with previous proposals for the esterification reaction

Author	***Feature Set***	***Idn***	***Type***
Orantes et al. [16]	[11, 26, 12, 22, 25, 1, 2, 3]	46.65	Filter
Uribe et al. [17]	[5, 14, 4, 6, 1, 2, 27, 3]	45.01	Wrapper
Uribe et al. [30]	Not Available	-	Wrapper
Expert-Guided SFS	[5, 14, 2, 1, 22, 18, 20, 15]]	40.48	Wrapper
Expert-Guided SFS	[21, 8, 6, 13, 19, 22, 3]	41.93	Wrapper

Conclusions and future work

An expert-guided wrapper for feature selection on data-based industrial process monitoring is presented. Expert knowledge is incorporated in the feature search to look for a subset of features able to represent the expert knowledge, but not in a supervised way, since it is important to take into account the data structure itself. Sequential Forward Selection (SFS) and Sequential Backward Elimination (SBE) were used as search methods, coupled with LAMDA as clustering algorithm and the Index of Dissimilarity to assess the cluster quality measure comparing the expert-knowledge partition with the clustering results.

The proposed methodology was successfully applied to a complex industrial process known as the Open Plate Reactor (OPR), on the thiosulfate and the esterification reaction. The objective was identify abnormal behaviors in the process when using relative simple sensor (temperature), even though some states concerns changes on flow composition of primary and secondary reactants. First, using a training data set, the subset of feature is selected and a behavioral model is constructed using just the reduced set of features. Then, the generated model was tested on a validation data set consisting of perturbations different than those used in training, including simultaneous faults. In both cases, the proposed approach was able to select a set of features capable of generating a behavioral model robust enough to identify not only all functional states on the train data set but correctly identify faults on the test dataset.

The proposed procedure was compared with previous approaches dealing with the same chemical reactions. A fewer number of features were needed to correctly identify all the functional states of the complex chemical process. The feature subset shows a good response and performance since the index of dissimilarity was lower than other approaches, indicating a high similarity with the expert-knowledge proposal. The main improvement of this methodology is introducing the unsupervised learning and expert guidance in the search process. The use of a non-iterative clustering algorithm leads to fast performance on the search over the feature subset space. Even though some specific methods were used at each block of the wrapper, the presented framework can be applied to any clustering method. Future work will consist in comparing different methods of feature selection, clustering, cluster quality and partition comparing to determine which among the methods proposed in the literature has better performance on specific applications.

Acknowledgment

Thanks to Dr. A. Orantes and DISCO Group at LAAS/CNRS for the access to the OPR databases. Thanks to CODI-Universidad de Antioquia and COLCIENCIAS for financial support.

References

1. F. Akbaryan, P. R. Bishnoi. "Fault diagnosis of multivariate systems using pattern recognition and multisensor data analysis technique". *Computers & Chemical Engineering*. Vol. 25. 2001. pp. 1313-1339.

2. I. Guyon, S. Gunn, M. Nikravesh, L. A. Zadeh. "Feature Extraction: Foundations and Applications" *Studies in Fuzziness and Soft Computing*. Vol. 207 Springer. Heidelberg. 2006. pp. 1-22.

3. D. W. Aha, R. L. Bankert. "A comparative evaluation of sequential feature selection algorithms". *In Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics*. Springer-Verlag. 1995. Fort Lauderdale. pp. 1-7.

4. K. R. Muske, C. Georgakis. "A methodology for optimal sensor selection in chemical processes". *Proc. American Control Conference the 2002*. Vol. 5. 2002. pp. 4274-4278.

5. R. Sikora, S. Piramuthu. "Efficient genetic algorithm based data mining using feature selection with hausdorff distance". *Inf. Tech. and Management*. Vol. 6. 2005. pp. 315-331.

6. L. M. Fraleigh, M. Guay, J. F. Forbes. "Sensor selection for model-based real-time optimization: relating design of experiments and design cost". *Journal of Process Control*. Vol. 13. no. 7. 2003. pp. 667-678.

7. S. Verron, T. Tiplica, A. Kobi. "Fault detection and identification with a new feature selection based on mutual information". *Journal of Process Control*. Vol. 18. 2008. pp. 479-490.

8. M. Bensch, M. Schroder, M. Bogdan, W. Rosenstiel. "Feature selection for high-dimensional industrial data". *Proceeding of the European Symposium of Artificial Neural Networks*. 2005. pp. 375-380.

9. T. Kourti. "Process analysis and abnormal situation detection: from theory to practice". *Control Systems Magazine*. IEEE. Vol. 22. no. 5. Oct 2002. pp. 10-25.

10. T. Kempowsky. "Surveillance de procédées à base de méthodes de classification". *Ph.D. dissertation*. INSA Toulouse. 2004. pp. 16-20.

11. P. Domingos. "The role of occam's razor in knowledge discovery". *Data Mining and Knowledge Discovery*. Vol. 3. 1999. pp. 409-425.

12. T.-H. Cheng, C.-P. Wei, V. Tseng. "Feature selection for medical data mining: Comparisons of expert judgment and automatic approaches". *Computer-Based Medical Systems*. 2006. pp. 165-170.

13. B. D. Burns, A. P. Danyluk. "Feature selection vs theory reformulation: A study of genetic refinement of knowledge-based neural networks". *Mach. Learn*. Vol. 38. 2000. pp. 89-107.

14. C. Isaza. "Diagnostic par techniques d'apprentissage floues : Conception d'une méthode de validation et d'optimisation des partitions". *Ph.D. dissertation*. Laboratoire d'Analyse et d'Architecture des Systèmes du CNRS. Toulouse. France. 2007. pp. 5-23.

15. L. E. Prat, A. Devatine, P. Cognet, M. Cabassud, C. Gourdon, S. Elgue, F. Chopard. "Performance evaluation of a novel concept "open plate reactor" applied to highly exothermic reactions". *Chemical Engineering and Technology*. Vol. 28. 2005. pp. 1028-1034.

16. A. Orantes, T. Kempowsky, M.-V. L. Lann, L. Prat, S. Elgue, C. Gourdon, M. Cabassud. "Selection of

sensors by a new methodology coupling a classification technique and entropy criteria". *Chemical Engineering Research and Design*. Vol. 85. no. 6. 2007. pp. 825-838.

17. C. Uribe, C. Isaza, O. Gualdron, C. Duran, A. Carvajal, "A wrapper approach based on clustering for sensors selection of industrial monitoring systems". *Broadband. Wireless Computing, Communication and Applications, International Conference on*. Vol. 1. 2010. pp. 482-487.

18. I. Guyon, A. Elisseeff. "An introduction to variable and feature selection". *J. Mach. Learn. Res*. Vol. 3. 2003. pp. 1157-1182.

19. S. Guerif, Y. Bennani. "Selection of clusters number and features subset during a two-levels clustering task" *Artificial Intelligence and Soft Computing*. 2006. pp. 28-33.

20. C. Isaza, A. Orantes, T. Kempowsky, M. Le Lann. "Contribution of fuzzy classification for the diagnosis of complex systems". *The 7th IFAC International Symposium of Fault Detection. Supervision and Safety of Technical Processes*. 2009. Barcelona. pp. 1132-1137.

21. T. Kempowsky, A. Subias, J. Aguilar-Martin. "Process situation assessment: From a fuzzy partition to a finite state machine". *Engineering Applications of Artificial Intelligence*. Vol. 19. no. 5. 2006. pp. 461-477.

22. J. Aguilar-Martin J., C. Isaza, E. Diez-Lledo, M.V. LeLann, J. Waissman-Vilanova. "Process Monitoring Using Residuals and Fuzzy Classification with Learning Capabilities". *Advances in Soft Computing* Springer Berlin Heidelberg. Volume 42. New York. 2007. pp. 275-284

23. C. Isaza, M.-V. L. Lann, J. Aguilar-Martin, "Diagnosis of chemical processes by fuzzy clustering methods: New optimization method of partitions". *18th European Symposium on Computer Aided Process Engineering (ESCAPE 10)*. 2008. pp. 1-6.

24. A. Orantes. "Methodologie pour le placement des capteurs a base de methodes de classification en vue du diagnostic". *Ph.D. dissertation*. Laboratoire d'Analyse et d'Architecture des Systemes du CNRS. 2005. pp. 29-39.

25. J. Aguilar-Martin and R. L. de Mantaras. "The process of classification and learning the meaning of linguistic descriptors of concepts". *Approximate Reasoning in Decision Analysis*. 1982. M.M. Gupta et E. Sanchez (eds.) North Holland. pp. 165-175.

26. J. Aguado, J. Aguilar-Martin. "A mixed qualitative-quantitative selflearning classification technique applied to diagnosis". *QR'99 The Thirteenth International Workshop on Qualitative Reasoning*. 1999. Loch Awe. pp. 124-128.

27. X. V. Nguyen, J. Epps., J. Bailey. "Information theoretic measures for clustering comparison: is a correction for chance necessary?". *ICML*. 2009. pp. 135.

28. R. Mantaras. "A distance-based attribute selection measure for decision tree induction". *Mach. Learn.*. Vol. 6. no. 1. 1991. pp. 81-92.

29. R. Mantaras. "Autoapprentissage d'une partition: application au classement iteratif de donnees multidimensionelles". *Ph.D. dissertation*. Univ. Paul Sabatier. Toulouse. 1979. pp. 20-37.

30. C. Uribe, C. Isaza. "Unsupervised feature selection based on fuzzy partition optimization for industrial processes monitoring". *Proccedings of the 2011 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications*. 2011. Ottawa. pp 1-5.