Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Physique 6 (2005) 1161–1168
COMPTES RENDUS PHYSIQUE
http://france.elsevier.com/direct/COMREN/

Physics/Applied physics

Detection of functional states by the 'LAMDA' classification technique: application to a coagulation process in drinking water treatment

Bouchra Lamrini [a,*], Marie-Véronique Le Lann [b,c], Ahmed Benhammou [a], El Khadir Lakhal [a]

[a] *Laboratoire d'automatique et d'étude des procédés, Faculty of Sciences Semlalia, PO Box 2390, 40000 Marrakech, Morocco*
[b] *LAAS-CNRS, 7, avenue du Colonel Roche, 31077 Toulouse cedex 4, France*
[c] *INSA, DGEI, 135, avenue de Rangueil, 31077 Toulouse cedex 4, France*

Received 1 July 2005; accepted 29 November 2005

Available online 6 January 2006

Presented by Jacques Villain

Abstract

The present Note proposes a learning classification methodology to identify functional states on a coagulation process involved in drinking water treatment. In this work, we chose to carry out the supervised control of this process while using the LAMDA (Learning Algorithm for Multivariate Data Analysis) classification technique. The LAMDA classification technique proposes the interactive participation of the expert operator during the learning phase and in the optimisation of the classification. In this work, all information stemming from the environment process as well as expert knowledge has been aggregated and exploited. The application chosen for state identification is the Rocade drinking water treatment plant located at Marrakech, Morocco. ***To cite this article: B. Lamrini et al., C. R. Physique 6 (2005).***

Résumé

Détection des états fonctionnels par la méthode de classification « LAMDA » : application à une unité de coagulation en traitement d'eau potable. Le travail présenté propose une méthodologie de classification par apprentissage qui permet l'identification des états fonctionnels sur une unité de coagulation impliquée dans le traitement d'eau potable. Nous avons choisi de réaliser la conduite supervisée de ce procédé en utilisant la méthode de classification LAMDA (Learning Algorithm for Multivariate Data Analysis). La technique de classification LAMDA permet une interaction directe de l'expert durant la phase d'apprentissage et d'optimisation de la classification. Dans ce travail, toutes les informations provenant du procédé lui même et de son environnement ainsi que les connaissances de l'expert ont été exploitées et agrégées. Le site d'application choisi pour l'identification des états fonctionnels est la station de traitement d'eau potable Rocade de Marrakech, Maroc. ***Pour citer cet article : B. Lamrini et al., C. R. Physique 6 (2005).***

Keywords: Classification; Learning; Pattern recognition; Fuzzy logic; Drinking water treatment process; LAMDA

* Corresponding author.
E-mail address: blamrini@yahoo.fr (B. Lamrini).

1631-0705/$ – see front matter
doi:10.1016/j.crhy.2005.11.017

Mots-clés : Classification ; Apprentissage ; Reconnaissance de forme ; Logique floue ; Traitement d'eau potable ; LAMDA

1. Introduction

The control and monitoring of drinking water treatment plants have become increasingly important and advanced world-wide. Indeed, in the case of complex processes, it is not always possible to derive a suitable mathematical or structural model. However, the use of other approaches as classification technique is needed to identify the functional state (normal or abnormal) and to decide whether a corrective action should be undertaken. The coagulation process is one of the most important stages in surface water treatment. It is strongly linked to the raw water quality. Monitoring of this process is essential for the maintenance of satisfactorily treated water quality and economic plant operation.

The main objective of this Note is the identification of functional states and the detection of faults on a coagulation unit from characteristics raw water. The use of learning classification techniques for the design of monitoring systems is becoming increasingly popular specially when dealing with this type of process. For this, we propose a methodology based on the LAMDA (Learning Algorithm for Multivariate Data Analysis) classification method which was conceived and studied within LAAS (Laboratory for Systems Analysis and Architecture) of CNRS, Toulouse. LAMDA methodology is a fuzzy generalisation of a probability function. Nevertheless, for concrete cases, it is equivalent to a statistical method. On the other hand, it can be a neural technique because it possesses a structure like that of Radial Basis Neural Networks. However, its theoretical bases are in the Fuzzy Logic domain, and all new developments of the method are carried out to keep coherent with the original motivations. Ref. [1] evoked different classification methods as the methods of strict and fuzzy clustering. Table 1 presents the general features of these classification methods in comparison to those of the LAMDA method.

Most essential is the fact of being able to propose the interactive participation of the expert during the learning phase and in the optimisation classes. It been shown that the user can adjust the different parameters which take part in the learning stage according to his experience and other criteria. By means of such an interaction between expert knowledge and the LAMDA algorithms, it is expected to improve the quality of the classification. So, the expert allows one to assign the different functional states to these classes.

This Note will first describe the application site chosen for state identification. A brief description of LAMDA classification technique is given in Section 3. The experimental results will then be presented and discussed.

2. Overview of study area

The drinking water treatment plant concerned in this study is the Rocade plant located at Marrakech, Morocco. It provides water to more than 1.5 millions inhabitants. Sixty percent of city needs are assured by this plant; the

Table 1
Principal features of classification methods

Tableau 1
Principales caractéristiques des méthodes de classification

	Linear discriminant	K-means clustering	GK-means clustering	Classification And Regression Tree (CART)	ANN Radial Basis Networks (RBN)	LAMDA
Parameter method	Yes	No	No	No	No	No
Data type	Numerical	Numerical	Numerical	Numerical & Qualitative	Numerical	Numerical & Qualitative
Pre-defined classes	Yes	Yes	No	Yes	Yes	Yes/No
Fixed number of predefined classes	Yes	Yes	Yes	Yes	No	No
Need to a known set of learning	Yes	Yes	Yes	Yes	Yes	Yes/No
Adjuster parameter	–	K [1]	K, m, ε [1]	–	Transfer function learning step	DAM Connectives Exigency Level [1]
Classification type	Strict	Strict	Fuzzy	Strict	Strict	Fuzzy & Strict
Classes update	No	No	No	No	No	Yes

complement is brought by underground resources (well, drilling, ...). It has a nominal capacity to process 1400 l/s of water. The treated water is stored in two tanks and transported toward the water supply network. The drinking treatment plant involves physical and chemical processes. The treatment consists, essentially, in the first disinfection, then coagulation-flocculation, settling, filtration and final disinfection.

3. LAMDA classification technique

The LAMDA (Learning Algorithm for Multivariate Data Analysis) methodology is a classification technique introduced by Aguilar-Martin [2] in the early 1980s and developed by Piera Carreté et al. [3]. It has been described successively in several works, particularly in [4–6] as a tool for the design of supervision and diagnosis systems of industrial processes. More recent studies, [7–9] have described in detail the methodology, as well as the algorithms and functions used. LAMDA is a fuzzy methodology of conceptual clustering and classification. It allows the representation of classes or concepts by means of the logic connection of all marginal information available. The formation and the recognition of classes are based on the attribution of each object to a class according to the heuristic rule of Maximal Adequacy. An object is then most likely to belong to the class which presents the Greater Adequacy Degree (*GAD*). It models the total indistinguishability (chaotic homogeneity) or homogeneity inside the description space from which the information is extracted. This is done by means of a special class called the Non Informative Class (*NIC*). This class accepts all items with the same adequacy; therefore it introduces naturally a classification threshold. The LAMDA classification technique allows also us to use different learning strategies such as:

- unsupervised learning: there is no previous knowledge of any class, there are no pre-defined classes, and the only existing class is the NIC;
- supervised learning: the user has previously defined a certain number of classes and has decided to which class each element must be assigned to;
- the third type of procedure is an extension of supervised learning, where some classes have been previously defined but it is possible to create other classes.

Other LAMDA characteristics are:

- it is able to use simultaneously numerical and qualitative information;
- learning is made in a sequential and incremental way;
- classification algorithms are based on linear compensated hybrid connectives which aggregates the marginal adequacies degrees (*MAD*) to obtain the global adequacy degree (*GAD*) of an object to a class;
- it is possible to obtain different classifications from the same group of objects by means of the 'exigency' concept.

Let us consider a collection of objects or individuals X, and a finite set of n qualitative or quantitative descriptors (attributes). For each class C_j, an object is associated to a vector $[M_{1,j}(x_1), \ldots, M_{n,j}(x_n)]$ where $M_{n,j}$ is a Marginal Adequacy Degree (*MAD*) of x_i. The information conveyed by each descriptor contributes to the membership of the element to the class by means of the Marginal Adequacy Degree. It must be noted that in order to calculate the adequacy of an element to a class, both must have the same descriptor set. Then, all the *MAD* are aggregated in order to obtain a Global Adequacy Degree (*GAD*) of the object to the class (Fig. 1). This is made by a convex interpolation of fuzzy logic connectives L_α called Mixed Connectives of linear compensation. The mixed connectives that interpolate between a conjunctive and a disjunctive logic operator are introduced and studied in [10]. It was shown that such interpolation is completely ordered with respect to the 'Exigency Level', the highest in the conjunction case (AND) and the lowest in the disjunction case. The following equation shows this concept:

$$GAD_\alpha(MAD_1, \ldots, MAD_n) = \alpha T(MAD_1, \ldots, MAD_n) + (1-\alpha)S(MAD_1, \ldots, MAD_n) \tag{1}$$

MAD_i is the marginal adequacy of the object and $\alpha \in [0, 1]$, to be coherent with Fuzzy Logic aspects, that include compatibility with Boolean Logic. T is an iterated t-norm and S its dual co-norm with respect to he negation (complement to 1). Parameter α is called the Exigency Index, and it is possible to associate different classifications to the same data set, depending on the value chosen for α. Recognition is more exigent [10] as α increases, therefore there will be more objects not recognized. Similarly if α increases, learning becomes more selective (or exigent) as the

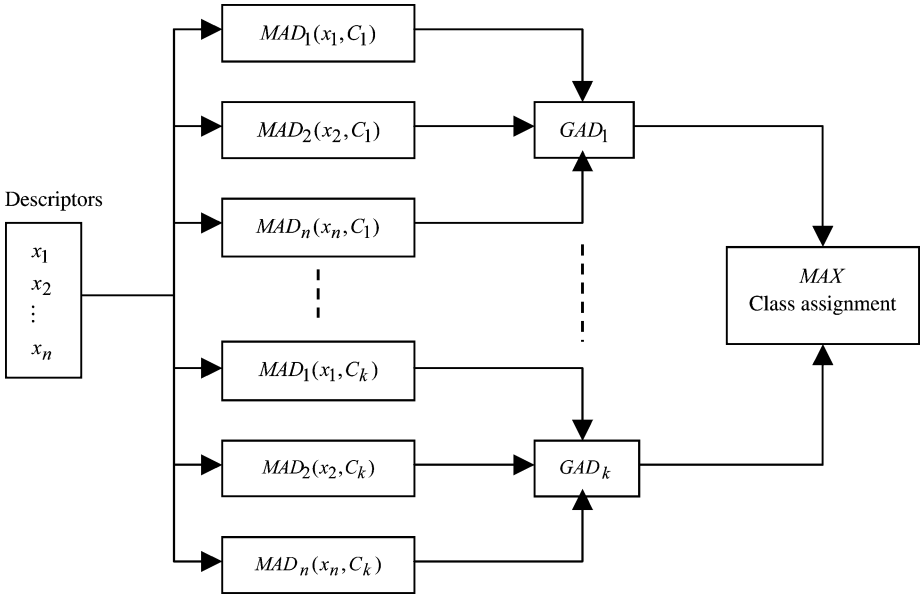

Fig. 1. Marginal and Global Adequacy Degrees.

Fig. 1. Degrés d'Adéquation Marginale et Globale.

number of objects assigned to the NIC increases, and so does the number of classes created. Thus, by changing the value of α, different partitions from the same data set, based on the same logical criteria, can be obtained.

To make the different calculations required in the learning and classification procedures it is necessary to choose the learning parameters such as:

- *Selection of the classification algorithm*: Given that *MAD* depends on the nature of each descriptor, the algorithm uses general possibility functions. For quantitative descriptors there are several options introduced in [11] to compute the *MAD*. One possibility function applied in this work is a fuzzy extension of the binomial probability function, which gives as result the following expression:

$$MAD_{c,d} = \rho_{c,d}^{X_{i,d}} (1 - \rho_{c,d})^{(1-X_{i,d})} \tag{2}$$

  where:
  - $\rho_{c,d}$ = Learning parameter for class c and descriptor d;
  - $X_{i,d}$ = Normalised value of the quantitative descriptor d for a particular individual i.
- *Selection of the connectives*: The user may choose between two different families of logic operators (T-norms and dual T-conorms functions) [1] in order to aggregate all the Marginal Adequacy Degrees of an individual to a class. These operators are: Product-Probabilistic Sum (T-norm and T-conorm) and the Minimum-Maximum (T-norm and T-conorm).
- *Selection of the exigency level*: The user may also use mixed connectives of the same family by choosing an exigency level (α) between 0 and 1. In recognition we call α the exigency level and in self-learning it is called selectivity level. By changing the value of α different partitions based on the same data used may be obtained. Thus, as the value of α increases, more classes will be created or in the case of recognition a greater adequacy is required of a measurement to the assigned to a pre-established class.
- *Selection of learning mode*: The user may use different learning such as self-learning (unsupervised learning) and supervised learning. In the self-learning, LAMDA creates a group of classes using as base the values of the descriptors from the individuals of the data file that has been loaded. There are two selectors that may be changed by the user only when unsupervised learning has been chosen. These are the maximum desired variation percentage and the maximum allowed iterations number. They are necessary because in unsupervised learning the class parameters vary, so to obtain some stability and to overcome in a certain way the effect of the observations ordering, the same data is classified several times until a maximum percentage of individuals changes from one

iteration to another. However, the time to reach stability could be very long, so a maximum number of iterations has also been introduced. This means that the data will be presented to the existing classes once and again until either no more than the specified percentage of individuals varies its assignment from the previous iteration or the cycles limit has been reached.

4. Results and discussion

Classification results presented in this section have been obtained with a self-learning mode. The experimental data of four years (from January 2000 to July 2003: nearly 1674 individuals) are used to identify the functional states. We used like describers the 5 characteristics of raw water Rocade plant such as: temperature (T), pH, TSS (Total Suspend Solids), dissolved oxygen (DO) and conductivity (COND). This data set, particularly temperature, pH and TSS, are strongly dependent on the seasonal phenomena (Fig. 2(a)).

In this work, algorithm that we chose to compute the marginal adequacy degrees is Lamda1 shown in (2), e.g., $MAD_{c,d} = \rho_{c,d}^{X_{i,d}}(1-\rho_{c,d})^{(1-X_{i,d})}$ function and the Minimum-Maximum was selected as the connective family. To calculate the global adequacy degrees, we adopted an Exigency Level equal of $\alpha = 0.8$. With the LAMDA software toolbox the process expert is able to increase the Exigency Level until he is satisfied with the resulting classes.

Fig. 2(b) shows the different classes obtained by the self-learning as well as the different functional states detected (see Table 3 too) on the coagulation process. These significant states are characterised from the classification information (class profile, membership matrix, etc.). While exploiting the information stem from profile classes (Fig. 3),

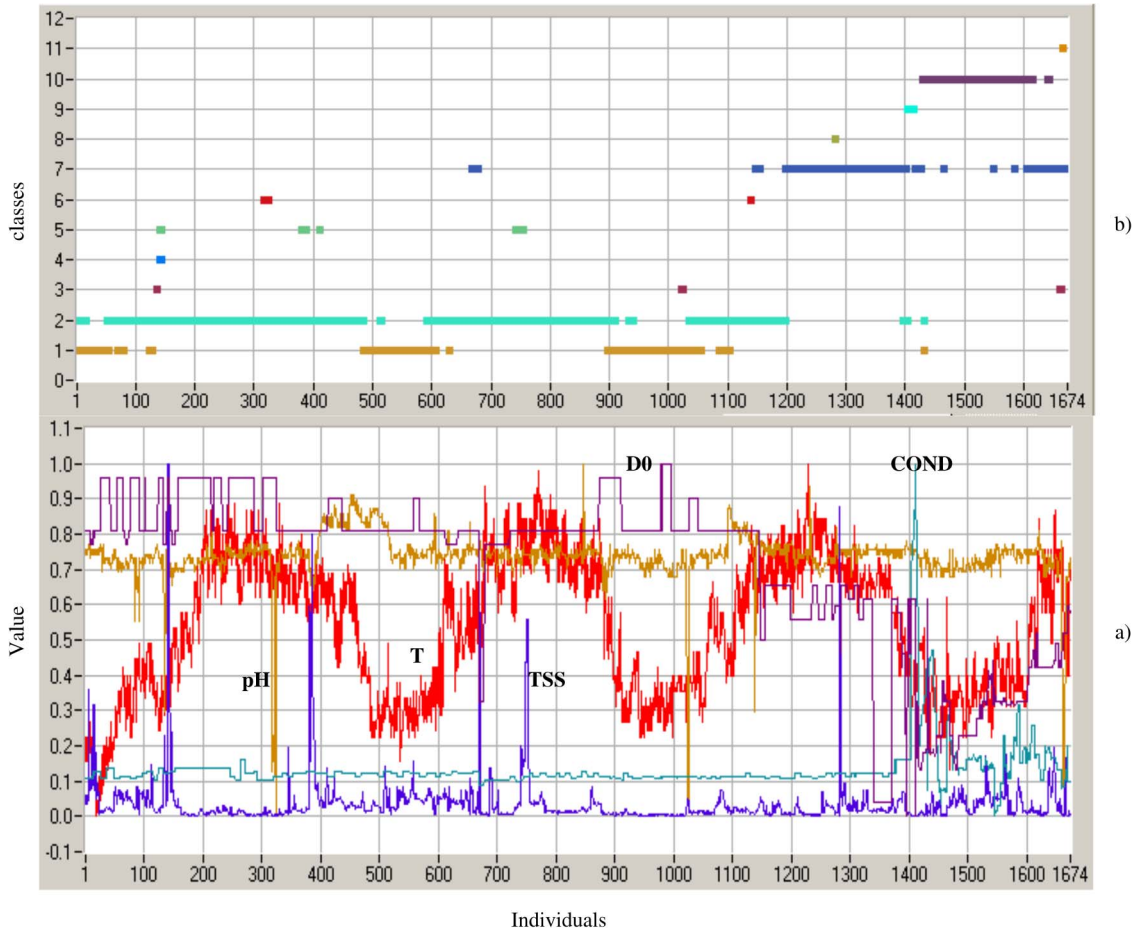

Fig. 2. (a) Normalised describers; (b) classification results for self-learning.

Fig. 2. (a) Descripteurs normalisés ; (b) classes issues d'auto-apprentissage.

Table 2
Describers frequency of every class

Tableau 2
Les valeurs normalisées des paramètres de chaque classe

	Number of elements	T-Freq	pH-Freq	TSS-Freq	COND-Freq	DO-Freq
Class 1	336	0.39337	0.74384	0.03889	0.11758	0.83674
Class 2	837	0.63469	0.75336	0.03987	0.11909	0.83340
Class 3	9	0.47444	0.14487	0.08514	0.13886	0.63885
Class 4	3	0.43333	0.67094	0.83333	0.34683	0.80769
Class 5	30	0.65133	0.70818	0.37569	0.13664	0.88463
Class 6	7	0.65350	0.30994	0.06677	0.15817	0.46551
Class 7	294	0.66188	0.73911	0.03317	0.13707	0.54808
Class 8	1	0.60556	0.61538	0.68976	0.31013	0.49113
Class 9	12	0.40085	0.73373	0.04158	0.75581	0.49113
Class 10	175	0.35346	0.73174	0.04163	0.17363	0.38606
Class 11	1	0.47333	0.58974	0.33333	0.31716	0.50963

Table 3
Functional states for self-learning

Tableau 3
Etats fonctionnels associés aux classes issues d'auto-apprentisage

Class	Class name	Associated state
1	LowSeason	Normal
2	HighSeason	Normal
3	pH_VeryLow_LowSeason∼∼TSS	pH Alarm
4	TSS_VeryBlevated	TSS(Slow→Stop)
5	TSS_Blevated	TSE Alarm
6	pH_VeryLow_HighSeason	pH Alarm
7	DO_Low_HighSeason	DO Alarm
8	TSS_VeryBlevated+DO_Low	(TSS_DO) Alarm
9	GOND_VeryBlevated∼∼DO	GOND Alarm
10	DO_VeryLow_LowSeason	DO Alarm
11	Describers- - - → Optimal_State	Transition

e.g., normalised parameters of every class, we can note that some classes present sometimes a similar characteristics and the expert can decide to regroup these classes in a single state. Eleven classes have been identified and according to their profile, eight functional states have been detected. Table 2 presents the parameters normalised (membership frequency of every describer) for every class. This information allows us to identify significant classes and those that can be regrouped in a single state.

- Class 1 and Class 2 profiles: the station operates in the normal conditions, e.g., the describers operate with the optimal values in the high and low season. We can therefore regroup these two classes in a single functional state called 'Normal'. 1153 elements have been associated to this state.
- Class 3 profile: in the low season, we clearly see an abnormal variation of pH (very low $pH_{freq} = 0.14$). A low variation of TSS ($TSS_{freq} = 0.08$) is noted simultaneously to this pH lowering. This class is associated to 'pHAlarm' state. 9 elements have been associated to this state.
- Class 4 profile: two very elevated variations of the TSS ($TSS_{freq} = 0.83$). The station is normally in a slowing state and it can change from this state to a stop state. The identified state is 'TSS (Slow- - → Stop)'. 2 elements have been associated to this state.
- Class 5 profile: in this case, we note high measures of TSS that are in the range ($0.12 < TSS < 0.8$). The Rocade plant operates in the normal conditions. Moreover, it is imperative to increase the coagulant dose. This class is associated to the 'TSSAlarm'. state. 20 elements have been associated to this state.

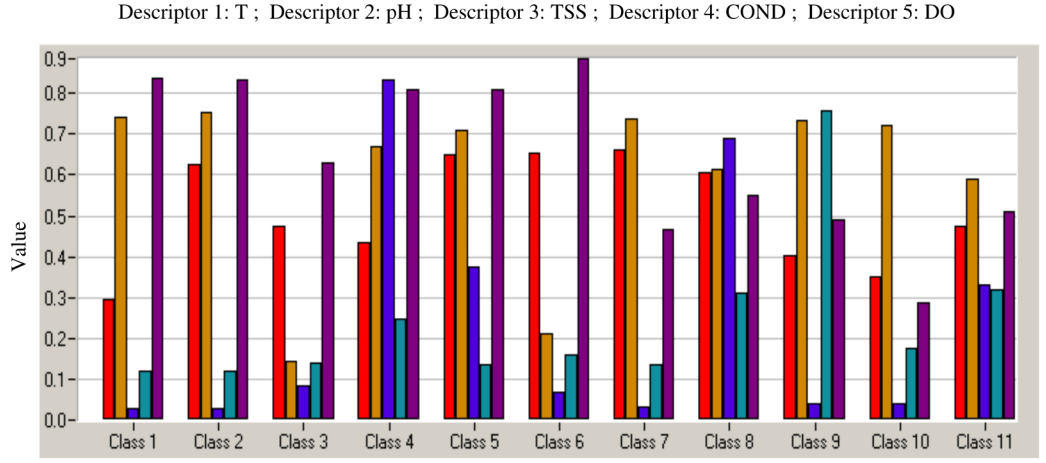

Fig. 3. Profile classes for self-learning.

Fig. 3. Profil des classes issues d'auto-apprentissage.

- Class 6 profile: an abnormal variation of pH (very low, $pH_{freq} = 0.20$) is noted in the high season. This class is associated to 'pHAlarm' state. 7 elements have been associated to this state.
- Class 7 profile: we note that the dissolved oxygen ($DO_{freq} = 0.46$) is decreased in the low season. The seventh class is associated to 'DOAlarm' state. 294 elements have been associated to this state.
- Class 8 profile: in the high season, TSS is very elevated ($TSS_{freq} = 0.68$) with a small variation of the dissolved oxygen (DO is in low state). The plant is in the '(TSS_DO)Alarm' state. Only one element has been associated to this state.
- Class 9 profile: the conductivity measures are very elevated (caused by the presence of chlorides) with a very low variation of dissolved oxygen. The state associated to this class is 'CONDAlarm' state. 12 elements have been associated to this state.
- Class 10 profile: abnormal variation of the dissolved oxygen in the high season (DO is low: $DO_{freq} = 0.28$) with a weak increase of the conductivity. The class 10 is associated to the 'DOAlarm' state. 175 elements have been associated to this state.
- Class 11 profile: the pH change from the low state to the normal state. This transition is noted simultaneously with transition of other describers toward the normal functioning state. We can associate this class to 'Transition' state. Only one element has been associated to this state.

5. Conclusion

In this work, we showed the performances of LAMDA classification method to identify the different functional states describing the behaviour coagulation process. A data set of four years stemming from Rocade treatment plant has been used. It was possible to identify 8 states of normal or abnormal functioning. This approach is a first application that shows the utility of classification techniques in the monitoring and the surveillance of this process type. It is clear that the final objective is to spread this monitoring to other treatment processes in order to detect at the earliest a drifts functioning or to identify a failures on an upstream unit.

References

[1] T. Kempowsky, Surveillance des procédés à base de méthodes de classification: conception d'un outil d'aide pour la détection et le diagnostic des défaillances, PhD Thesis, LAAS-CNRS, Institut National Polytechnique de Toulouse, France (2004).

[2] J. Aguilar-Martin, M. Balssa, R. Lopez De Mantras, Estimation récursive d'une partition : Exemples d'apprentissage et auto-apprentissage dans R^n et I^n, Rapport technique 880139, LAAS-CNRS, Toulouse, France (1980).

[3] N. Piera Carreté, P. Desroches, J. Aguilar-Martin, LAMDA: An incremental conceptual clustering system, Rapport technique 89420, LAAS-CNRS, Toulouse, France (1989).

[4] J. Aguilar-Martin, R. López de Mántaras, The process of classification and learning the meaning of linguistic descriptors of concepts, in: Approximate Reasoning in Decision Analysis, North-Holland, Amsterdam, 1982, pp. 165–175.
[5] M. Chan, J. Aguilar-Martin, N. Piera Carreté, P. Celsis, J.P. Marc-Vergnes, Classification techniques for feature extraction in low resolution tomographic evolutive images: Application to cerebral blood flow estimation, in: 12th Conference GRESTI, 1989.
[6] N. Piera Carreté, P. Desroches, J. Aguilar-Martin, Variation points in pattern recognition, Pattern Recognition Lett. 11 (1990) 519–524.
[7] J.C. Aguado, A mixed qualitative-quantitative self-learning classification technique applied to situation assessment in industrial process control, PhD Thesis, Universitat Politècnica de Catalunya (1998).
[8] J. Waissman, J. Aguilar-Martin, B. Dahhou, G. Roux, Généralisation du degree d'adéquation marginale de la méthode de classification LAMDA, 6$^{\text{ièmes}}$ rencontres de la Société Francophone de Classification (1998).
[9] J. Waissmann, Construction d'un modèle comportemental pour la supervision de procédés : Application a une station de traitement des eaux, PhD Thesis, LAAS-CNRS, Institut National Polytechnique de Toulouse, France (2000).
[10] N. Piera Carreté, J. Aguilar-Martin, Controlling selectivity in non-standard pattern recognition algorithms, IEEE Trans. in Syst. Man and Cybernetics 21 (1991) 71–82.
[11] J. Waissman-Vilanova, J. Aguilar-Martin, B. Dahhou, G. Roux, Généralisation du degré d'adéquation marginale dans la méthode de classification LAMDA, Report 98396, LAAS-CNRS, Toulouse, France (1998).