Evaluation de la pertinence de paramètres biochimiques et classification pour la caractérisation des états physiologiques dans un bio-procédé par la théorie de l'évidence

Sébastien Régis[*,***], Andrei Doncescu[*]
Jean-Pierre Asselin de Beauville[**,+], Jacky Desachy[***]
[*]LAAS-CNRS/ LBB INSA 7, av. du col. Roche 31077 Toulouse Cedex 04
adoncesc@laas.fr
http://www.laas.fr
[**]Laboratoire d'Informatique de l'Université de Tours 64 av. J. Portalis 37200 Tours
jean-pierre.asselin@auf.org
http://www.e3i.univ-tours.fr
[+] en détachement à l'Agence Universitaire de la Francophonie, Montréal Canada
[***]GRIMAAG Université Antilles-Guyane Campus de Fouillole 97159 Pointe-à-Pitre
jdesachy, sregis@univ-ag.fr
http://www.univ-ag.fr/grimaag/

Résumé. L'analyse et la modélisation des bio-procédés nécessitent une connaissance profonde des systèmes biologiques. Ces phénomènes biologiques sont particulièrement complexes et ne peuvent être modélisés totalement, même par un système différentiel non linéaire. Le but de ces modélisations mathématiques est la détection des états physiologiques. Il est aussi possible de chercher à détecter les états physiologiques en utilisant les signaux biochimiques mesurés en ligne. Les signaux utilisés pour la classification sont souvent peu nombreux. Nous présentons une méthode de classification des paramètres biochimiques basée sur la théorie de l'évidence. Cette théorie est aussi utilisée pour évaluer la pertinence des paramètres. Cette évaluation est basée sur la notion de conflit. Nous présentons une alternative à la mesure classique de conflit; cette nouvelle mesure du conflit basée sur une distance, fournit des résultats plus cohérents pour l'application présentée. Les premiers résultats concernant l'analyse des états physiologiques d'un procédé biotechnologique de fermentation sont présentés.

1 Introduction

Durant ces dernières décennies, la biologie a connu un développement prolifique dans toutes ses facettes. A l'instar de la physique au siècle dernier, la biologie fournit de nouveaux défis et champs d'études aux mathématiques, à l'informatique et aux technologies. De nouveaux champs d'application et d'études comme la bio-informatique ou les biotechnologies ont pris aujourd'hui une place prépondérante dans la recherche fondamentale ou appliquée. Ainsi, les bio-procédés industriels ou expérimentaux qui utilisent des micro-organismes, font appel à de nombreux champs transversaux aux mathématiques et à l'informatique. La classification des paramètres biochimiques en

particulier, mesurés pendant le bio-procédé permet de détecter des états physiologiques nouveaux ou connus. Par ailleurs, les experts microbiologistes tentent également de trouver quels sont les paramètres biochimiques les plus pertinents. Un paramètre biochimique est dit pertinent s'il fournit une information significative pour trouver les états physiologiques. Cependant l'évaluation des paramètres pertinents n'est pas chose facile puisqu'elle est basée essentiellement sur des connaissances *empiriques* et *subjectives*. Le *problème* des microbiologistes peut être ainsi résumé : (1) trouver une classification des paramètres biochimiques qui corresponde aux états physiologiques (un état correspond à une ou plusieurs classes),(2) trouver les paramètres biochimiques pertinents.
Dans ce papier, nous présentons une méthode de classification basée sur la fusion d'information provenant de plusieurs paramètres biochimiques mesurés durant un procédé de fermentation. La classification par fusion de données est principalement basée sur la théorie de Dempster-Shafer (DS) ou théorie de l'évidence. Mais cette théorie n'est pas seulement utilisée pour effectuer la classification; elle permet également de trouver les paramètres les plus pertinents grâce à la notion de conflit. L'originalité de cette approche est donc l'utilisation de la théorie de l'évidence pour évaluer la pertinence des sources d'information. Il s'agit d'une nouvelle application possible de la théorie de l'évidence.
Le papier est organisé comme suit. Dans le paragraphe 2, nous présentons le problème biologique. Dans le paragraphe 3 est présentée la méthode LAMDA (Learning Algorithm for Multivariate Data Analysis) [Aguilar-Martin *et al.*, 1980] qui fournit une classification préliminaire permettant de calculer les masses d'évidence nécessaires à la théorie de l'évidence. Dans le paragraphe 4, nous présentons un rappel sur la théorie de l'évidence et nous expliquons le lien entre la pertinence des paramètres et la notion de conflit. Une alternative à la notion classique de conflit est proposée; cette nouvelle mesure du conflit basée sur une distance fournit des résultats intéressants mais n'est utilisable que dans certains cas précis. Cette notion de conflit basée sur la distance fournit des résultats plus cohérents que la notion de conflit classique dans l'application biotechnologique présentée. Enfin, dans le paragraphe 5 une première analyse des résultats expérimentaux est présentée.
Notation.
Pour chaque paramètre une mesure est effectuée à l'instant donné t. Chaque paramètre est donc un signal discrètisé en fonction du temps.
$x_i(t)$ est la mesure du paramètre i à l'instant t. Quand il n'y a pas d'ambiguité, on notera x_i au lieu de $x_i(t)$. x_i est alors appelé *échantillon* ou *élément* du paramètre i.
On dénote $x(t) = (x_1(t),x_2(t),...,x_n(t))$ les mesures des paramètres 1,2,...,n à un instant donné t. Quand il n'y a pas d'ambiguité, on dénote x.
La classe j est notée C_j.
c_j est le centre de la classe j. c_j est un vecteur de dimension n (où n est le nombre de paramètres). $c_{i,j}$ est la ième composante du centre de la classe C_j.

2 Les bio-procédés

Les procédés biotechnologiques font appel à des connaissances pointues sur la physiologie cellulaire, voire sur les connaissances les plus récentes en matière d'études

génomiques. Les méthodes mathématiques et informatiques utilisées dans ces bio-procédés ont pour but principal de mieux détecter (et donc mieux comprendre) les états physiologiques des micro-organismes utilisés afin de mieux optimiser le processus en fonction de l'objectif industriel ou expérimental. Parmi ces méthodes on distingue principalement :

- les modèles mathématiques. Il s'agit de trouver un modèle susceptible de reconstituer les phases de croissance des micro-organismes. La difficulté d'utilisation de ces modèles vient de la complexité du vivant, en effet, il est souvent nécessaire de multiplier les paramètres dans l'équation qui définit le modèle. Par ailleurs, les modèles varient en fonction des micro-organismes utilisés (nous avons dénombré plus de 130 modèles) et ceux-ci ne sont pas toujours en adéquation avec les mesures expérimentales, soit en raison d'une anomalie quelconque lors du bio-procédé, soit parce que le modèle n'est pas adapté à la situation.
- les techniques issues de l'intelligence artificielle. Elles cherchent à modéliser de façon explicite les connaissances des experts (voir par exemple [Steyer, 1991]). Cependant le nombre de règles des experts peut augmenter de façon quasi exponentielle et la modélisation de ces connaissances sous forme de règles pour un système expert est un travail long et fastidieux. D'autre part, la communauté des microbiologistes, face aux récentes découvertes, a tendance à remettre en question ses propres connaissances sur les lois du vivant, ce qui pourrait remettre en question ce type d'approche utilisant les systèmes experts.
- les techniques issues de la classification. On cherche à regrouper les éléments des mesures effectuées en ligne dans des classes de telle sorte que ces classes soient bien différentes les unes des autres tout en ayant pour chacune d'entre elles, la plus grande homogènéité possible. Ces techniques sont intéressantes en ce sens où elles peuvent faire appel ou non aux connaissances des experts

Bien que la tendance actuelle soit de chercher à fusionner ces trois approches, nous nous intéresserons plus spécifiquement à la dernière méthode citée qui utilise les techniques de classification. Nous avons déjà montré d'ailleurs que la méthode LAMDA fournit de bons résultats dans le domaine des bio-procédés [Regis *et al.*, 2003]. Cette méthode fournit une classification pour chacun des paramètres qui sert au calcul des masses d'évidence.

3 Classification préliminaire par la méthode LAMDA pour les masses d'évidence

La théorie de l'évidence nécessite l'utilisation de masses d'évidence. LAMDA fournit des degrés d'appartenance à des classes qui après modification (voir équation 3), correspondent à ces masses d'évidence. LAMDA est une méthode de classification qui peut être supervisée ou non supervisée, développée au LAAS de Toulouse qui tente de concilier les propriétés de la loi bayésienne et celles des méthodes neuronales simplifiées, tout en utilisant des opérateurs d'aggrégation flous issus de l'intelligence artificielle (voir [Piera-Carreté et Aguilar-Martin, 1991], [Waissman-Vilanova *et al.*, 1998] ainsi que [Nakkabi *et al.*, 2002]). La méthode LAMDA a été utilisée en reconnaissance

de forme [Piera-Carreté *et al.*, 1990], en analyse biomédicale [Chan *et al.*, 1989], pour l'étude des processus de dépollution des eaux usées [Waissman-Vilanova *et al.*, 2000] et pour les bio-procédés [Aguilar-Martin *et al.*, 1999]. Pour notre application, nous utilisons la version non supervisée de LAMDA de telle sorte que les experts microbiologistes puissent valider ou non *a posteriori* la classification. En effet, nous avons constaté qu'en mode non supervisé, LAMDA fournissait de bons résultats [Regis *et al.*, 2003]. Pour chaque paramètre biochimique, LAMDA calcule un degré d'appartenance associé à chacune des classes existantes (ces classes sont créées au fur et à mesure de la classification [Nakkabi *et al.*, 2002]) en utilisant une généralisation d'une loi binomiale appelée Degré d'Adéquation Marginal (DAM). Ce Degré d'Adéquation Marginal, proposé par Aguilar-Martin [Aguilar-Martin *et al.*, 1980], puis modifié par Waissman-Vilanova [Waissman-Vilanova, 2000], est défini par l'équation suivante :

$$DAM_{j,i}(x_i) = \rho_{ji}^{1-\alpha(x_i,c_{j,i})}(1-\rho_{ji})^{\alpha(x_i,c_{j,i})} \quad (1)$$

où $\rho_{i,j}$ est la probabilité qu'un élément appartienne à la classe C_j et $\alpha(x_i,c_{i,j})$ représente la distance normalisée entre x_i et $c_{i,j}$.
Le nombre de classes et les classes elles-mêmes ne sont pas connus à l'avance. Les éléments sont traités de façon séquentielle : les classes sont modifiées au fur et à mesure que les échantillons sont introduits dans la classification et de nouvelles classes sont éventuellement créées. Afin de savoir s'il faut ou non créer de nouvelles classes, une première fusion de l'information issue de toutes les sources (c'est-à-dire de tous les paramètres biochimiques) est réalisée en utilisant un opérateur d'aggrégation. Il existe divers opérateurs d'aggrégation (T-norme et T-conorme, moyenne, etc) mais LAMDA utilise le triple Π développé par Yager et Rybalov [Yager et Rybalov, 1998] pour sa propriété de renforcement total. Le triple Π est utilisé pour calculer le Degré d'Adéquation Global (DAG) pour chaque classe j :

$$DAG_j(x) = \frac{1}{1+\Pi_{i=1}^{n}\left[\frac{1-DAM_{j,i}(x_i)}{DAM_{j,i}(x_i)}\right]} \quad (2)$$

Si quelle que soit la classe j, le DAG est inférieur à 0,5, alors une nouvelle classe est créée (et x est le centre de cette nouvelle classe) et une nouvelle évaluation de tous les DAM est effectuée pour toutes les classes. On pourrait se demander à ce stade pourquoi l'on tient compte de l'information issue de tous les paramètres biochimiques alors que l'on cherche à garder uniquement ceux qui sont pertinents, mais cette première fusion présente un avantage certain. En effet, à partir de cette information globale on peut déterminer s'il faut créer une nouvelle classe ou non. De ce fait on tient compte de toutes les classes possibles: on passe ainsi d'un *monde ouvert* à un *monde fermé*. Un monde est considéré comme *fermé* si toutes les hypothèses possibles sont utilisées pour décrire les évènements de ce monde ; sinon ce monde est *ouvert*. Dans notre cas, un monde est ouvert si certains éléments de la classification forment un groupe qui ne ressemble à aucune des classes existantes (la classe correspondante à ce groupe a pu être omise); ce groupe d'éléments sera très mal classé ou ne sera pas classé du tout. On peut se retrouver dans ce cas si l'on doit fixer le nombre de classes *a priori*, et que ce nombre de classes fixé est inférieur au nombre réel de classes. Cette notion de monde

ouvert ou fermé est très importante pour la théorie de DS car elle peut grandement influencer la classification finale.
Nous pouvons maintenant faire deux remarques.
Premièrement, LAMDA fournit des degrés d'appartenance à des classes; une normalisation est donc nécessaire pour obtenir des masses d'évidence dont la somme soit égale à 1 (voir l'équation 3).
Deuxièment, tous les paramètres biochimiques travaillent exactement sur les mêmes classes (et pas seulement sur le même monde). Cette remarque est importante pour comprendre pourquoi nous proposons une alternative à la notion classique de conflit de la théorie de l'évidence.

4 La théorie de Dempster-Shafer et son application

4.1 Théorie de l'évidence

La théorie de l'évidence est une généralisation de la théorie bayésienne qui tient compte des notions d'incertitude et d'imprécision de l'information. Elle a été introduite par Dempster [Dempster, 1968] puis a été formalisée mathématiquement par Shafer [Shafer, 1976]. Considérons l'ensemble de toutes les évènements possibles (on parle d'ensemble de toutes les hypothèses) ; cet ensemble est appelé *ensemble de discernement* et est noté Θ. Toutes ces hypothèses sont mutuellement exclusives et sont nommées *singletons*. La théorie de Dempster-Shafer porte sur l'ensemble des sous-ensembles A de Θ. Cet ensemble de sous-ensembles de Θ est noté 2^Θ. A peut être composée d'un singleton ou d'une union de plusieurs singletons. Une fonction de masse $m(.)$ peut être alors définie de 2^Θ vers [0,1] avec les propriétés suivantes :

$$\sum_{A \subset \Theta} m(A) = 1 \qquad (3)$$
$$m(\oslash) = 0$$

$m(A)$ est la masse d'évidence associée à A. Les fonctions de *plausibilité* ($Pl(.)$) et de *croyance*($Bel(.)$) sont définies de 2^Θ vers [0,1] comme suit :

$$Pl(A) = \sum_{B \cap A \neq \oslash} m(B) \qquad (4)$$
$$Bel(A) = \sum_{B \subset A} m(B)$$

Pour obtenir une fusion de l'information de deux sources différentes 1 et 2, il existe une combinaison de leurs masses d'évidence appelée règle de Dempster-Shafer :

$$(m_1 \oplus m_2)(A) = m_{1,2}(A) = \frac{1}{1-K} \sum_{B \cap C = A} m_1(B).m_2(C) \quad A,B,C \subset 2^\Theta \qquad (5)$$

où K est défini comme suit :

$$K = \sum_{B \cap C = \oslash} m_1(B).m_2(C) \qquad (6)$$

Le dénominateur $1-K$ est un facteur de normalisation. Plus précisément K représente la mesure du conflit entre les sources 1 et 2. Plus K est important, plus les sources sont en conflit et moins la fusion a de sens. Si $K = 1$ alors le conflit est total et la fusion n'a pas de sens. On peut généraliser la règle de Dempster-Shafer à n sources :

$$(\oplus m_i)_{i=1,\ldots,n}(A) = \frac{1}{1-K} \sum_{X_1 \cap \ldots \cap X_n = A} \left(\Pi_{i=1}^{n} m_i(X_i)\right) \quad A, X_i \subset 2^{\Theta} \tag{7}$$

$$K = \sum_{X_1 \cap \ldots \cap X_n = \emptyset} \left(\Pi_{i=1}^{n} m_i(X_i)\right)$$

Si les sources sont en conflit fort (K est grand) alors la règle de DS peut conduire à des résultats erronés, en particulier si l'on travaille dans un monde ouvert. Dans ce cas, la bonne hypothèse a dû être omise [Zadeh, 1984][Smets, 1990]. Cependant comme nous travaillons dans un monde fermé, s'il y a conflit entre les classes, cela provient du fait qu'au moins une des sources est erronée ou non pertinente. Les sources doivent donc être suffisamment en accord pour que la fusion ait un sens.
Une fois que la fusion de l'information est réalisée il est possible d'utiliser différentes règles de décision pour effectuer la classification. Le choix du maximum de plausibilité correspond à un choix optimiste tandis que le choix du maximum de croyance correspond à une décision pessimiste. Le choix d'une règle de décision dépend à la fois de l'application et du comportement souhaité.

4.2 Pertinence des paramètres biochimiques et notion de conflit dans la théorie de l'évidence

Comme nous l'avons dit plus haut, les experts en microbiologie cherchent non seulement à mieux connaître les états physiologiques mais aussi à caractériser la pertinence des paramètres biochimiques. Rappelons que la pertinence des paramètres est basée sur des connaissances *subjectives* et *empiriques*. Bien sûr ces connaissances sont à la base fondées sur des connaissances biologiques et biophysiques, mais le nombre des connaissances mises en jeu lors d'un bio-procédé étant très important, seul l'expérience du microbiologiste lui permet de faire un tri de celles-ci pour comprendre les phénomènes expérimentaux, évaluer la pertinence des paramètres lors des ces expériences, et éventuellement prendre une décision si cela est nécessaire. De plus les connaissances de ces experts concernent surtout 5 ou 6 paramètres alors qu'il peut y en avoir beaucoup plus (entre 6 et 70, voire plus). De ce fait, une partie des informations fournies par l'ensemble des paramètres peut être soit inexploitée, soit redondante, voire erronée. Ainsi, en plus d'avoir une classification automatique des données qui leur fournit les états physiologiques, les microbiologistes désirent une évaluation de la pertinence des paramètres biochimiques ayant une base *objective* ou du moins suffisamment *théorique*. Par conséquent une méthode probabiliste qui suppose l'indépendance des paramètres, comme la classification bayesienne naïve, ne serait pas adaptée quand bien même elle fournirait une classification correcte, car elle ne donnerait aucune information sur la pertinence de ces paramètres. C'est ici que la théorie de l'évidence peut fournir une aide pour l'évaluation de la pertinence de ces paramètres. Nous proposons

d'utiliser la notion de conflit pour caractériser la pertinence des paramètres biochimiques.
En effet, en calculant la pertinence deux à deux entre les paramètres, il est possible de savoir quels sont les paramètres en accord et quels sont ceux qui sont en conflit. Si un paramètre est en conflit avec la majorité des autres paramètres (c'est-à-dire avec plus de la moitié des autres paramètres), celui est considéré comme non pertinent, sinon il est considéré comme valide. Il est ainsi possible de caractériser la pertinence d'un paramètre avec une certaine flexibilité. Cette caractérisation est faite pour chaque échantillon; la pertinence est donc évaluée de façon locale. Cette caractérisation locale est *a priori* plus significative qu'une caractérisation globale qui ne tiendrait pas compte des évolutions possibles au cours du temps. Il est alors possible d'enlever localement de la classification de Dempster-Shafer les sources qui ne sont pas pertinentes. Cette notion de pertinence utilisant la théorie de l'évidence peut être rapprochée de la fusion par vote de Dubois et Prade [Dubois et Prade, 1992] qui utilise la notion de nombre (fixé *a priori*) de sources valides; cependant pour autant que nous le sachions, dans cette méthode de fusion par vote, il n'y a pas de caractérisation des sources (i.e. des paramètres). Les méthodes statistiques traditionnelles comme l'Analyse en Composantes Principales (ACP) pourraient être aussi utilisées pour évaluer la pertinence des signaux. En effet, on peut considérer que x est un point de l'espace R^n où n est le nombre de paramètres, et tester ces méthodes sur le nuage de points de R^n. Cependant remarquons que ces méthodes fournissent en général une information globale et non locale. Ainsi on teste une ACP sur l'ensemble des points du début à la fin de l'expérience. On pourrait effectuer cette ACP sur des intervalles de temps dont les bornes sont définies à partir des singularités des paramètres comme cela est proposé dans [Regis *et al.*, 2004b], mais le début ou la fin d'un conflit ne correspond pas forcément à une singularité (voir la section 5). D'autre part, ces méthodes cherchent les corrélations entre les paramètres, or si le fait que les signaux soient fortement corrélés peut éviter la redondance d'information, deux signaux décorrélés ne sont pas forcément en conflit. Ainsi ces méthodes statistiques ne semblent pas adaptées à la recherche du conflit entre paramètres.

4.3 Vers une autre mesure du conflit

Comme on l'a vu, la valeur K permet de mesurer le conflit entre sources. Néanmoins, le conflit doit être suffisamment faible pour que la fusion ait un sens. Pourtant en utilisant la mesure de conflit K on peut arriver à des résultats erronnés. Ainsi pour deux paramètres fournissant exactement les mêmes masses d'évidence, on peut trouver un conflit K non nul. Prenons deux exemples. Premièrement supposons que nous avons 3 paramètres P1, P2 et P3 et deux classes C_1 et C_2. Les masses d'évidence supposées connues pour les 2 classes sont respectivement pour P1, P2 et P3 : $m_1(C_1) = \frac{1}{3}$, $m_1(C_2) = \frac{2}{3}$; $m_2(C_1) = \frac{1}{3}$, $m_2(C_2) = \frac{2}{3}$; $m_3(C_1) = 0$ et $m_3(C_2) = 1$. Nous pouvons calculer le conflit $K_{1,2}$ entre P1 et P2, et $K_{1,3}$ entre P1 et P3. La fusion d'information se fait comme indiqué dans le tableau 1 (les ensembles vides représentent le conflit). Le calcul du conflit de l'équation 6 se simplifie et revient dans ce cas à la formule suivante :

$$K_{i,j,i\neq j} = \sum_{k,l,k\neq l} m_i(C_k).m_j(C_l) \tag{8}$$

	$m_2(C_1)$	$m_2(C_2)$
$m_1(C_1)$	$m_{1,2}(C_1)$	$\emptyset$
$m_1(C_2)$	$\emptyset$	$m_{1,2}(C_2)$

TAB. 1 – *Fusion des masses d'évidence de P1 et P2*

	Masses d'évidence de T°
$m(C_1)$	0.111612
$m(C_2)$	0.102348
$m(C_3)$	0.101244
$m(C_4)$	0.112802
$m(C_5)$	0.151594
$m(C_6)$	0.107448
$m(C_7)$	0.107448
$m(C_8)$	0.107264
$m(C_9)$	0.098240

TAB. 2 – *Les masses d'évidence de la température*

Le conflit $K_{1,2}$ est égal à 0.444 alors que les paramètres P1 et P2 sont parfaitement en accord. De plus le conflit $K_{1,3}$ est égal à 0.333. Ainsi le conflit entre deux paramètres fournissant exactement les mêmes masses d'évidence est non nul et même supérieur au conflit entre deux paramètres donnant des résultats différents.
Prenons un exemple particulièrement parlant tiré de notre application. Considérons le paramètre Température (T°) qui fournit (comme tous les autres paramètres) des masses d'évidence sur 9 classes. Ces masses d'évidence sont présentées dans le tableau 2. Supposons que nous ayons un autre paramètre P4 qui fournisse exactement les mêmes masses d'évidence que celles de T° présentées dans le tableau 2. Et bien le conflit K entre T° et P4 serait égal à 0.886865 ! Ainsi deux paramètres fournissant la même classification seraient considérés comme étant en conflit fort ! Ceci est dû au fait que K est adapté à la mesure de conflit de sources travaillant sur des unions de classes différentes. Or ici, tous les paramètres travaillent sur les mêmes classes. Les masses d'évidences sont calculées pour des singletons (un singleton est ici une classe) et non sur des unions de singletons. En effet les masses d'évidences sont calculées en effectuant une normalisation; elles sont donc équivalentes à des probabilités. Or comme le notent Dubois et Prade [Dubois et Prade, 2004], bien que la théorie des probabilités soit un cas particulier de la théorie de l'évidence, cette dernière est rarement utilisée pour la fusion des probabilités fournies par diverses sources, en raison de la définition du conflit K (équations 6 et 8).
Ainsi on est face à une situation paradoxale : comme nous l'avons vu dans le paragraphe 4.1, l'utilisation de la théorie de Dempster-Shafer implique que le conflit entre sources soit faible, et pourtant dans ce cas particulier, la théorie de Dempster-Shafer ne détecte pas les sources qui n'ont aucun conflit. La notion classique de conflit ne semble donc pas adaptée à cette application. En fait, pour les exemples présentés ci-dessus, une approche

intuitive pour mesurer le conflit entre paramètres serait de calculer la distance entre les masses d'évidence de chaque classe. Cette approche conduit à la définition d'une nouvelle mesure de conflit basée sur la norme 1. Nous rappelons la définition de la norme 1 pour $y \in R_n$:

$$||y||_1 = \sum_{i=1}^{n} |y_i| \tag{9}$$

La nouvelle mesure du conflit que nous proposons est définie comme suit :

$$D = \frac{1}{2} \sum_{i} |m1(C_i) - m2(C_i)| \tag{10}$$

où $m1(C_i)$ et $m2(C_i)$ représentent les masses d'évidence pour des paramètres 1 et 2 pour la classe i. Le facteur $\frac{1}{2}$ est un facteur de normalisation. On peut noter que la norme est mathématiquement équivalente à la norme Euclidienne et à la norme Sup sur R^n (avec n fini), il est donc possible de définir une mesure de conflit basée sur l'une de ces deux autres normes. Si nous reprenons les deux exemples précédents, le calcul du conflit devient plus cohérent. Ainsi le conflit $D_{1,2}$ entre P1 et P2 est égal à 0 et le conflit $D_{1,3}$ entre P1 et P3 est égal à 0.333. Pour l'exemple tiré du bio-procédé on trouve maintenant que le conflit entre T° et P3 est nul. De plus il n'est pas nécessaire de définir une mesure du conflit pour n paramètres puisque l'on calcule le conflit deux à deux entre les paramètres.
Il est clair que ce conflit D peut être utilisé uniquement dans le cas où les masses d'évidence sont calculées pour des singletons et non des unions de singletons. Dans ce dernier cas, on utilisera la mesure du conflit classique K pour évaluer la pertinence des paramètres. Si l'on n'a aucune information sur les masses d'évidence c'est-à-dire si l'on ne sait pas si ces masses d'évidence sont calculées pour des singletons ou pour des unions de singletons, il est préférable d'utiliser la mesure de conflit classique K. Sinon la mesure D risque de fournir des résultats eronnés si l'on travaille sur des unions de singletons (par exemple pour deux unions de singletons dont l'intersection est vide, la mesure D donne un conflit nul alors que le conflit n'est pas nul). Il faut cependant se rappeler que l'idée de base de cette approche est moins l'utilisation d'une nouvelle mesure du conflit que l'utilisation de la théorie de l'évidence pour caractériser la pertinence des sources d'informations.

5 Résultats expérimentaux

Le bio-procédé que nous étudions est un bio-procédé de fermentation utilisant les micro-organismes appelés *Saccharomyces Cerevisiae*. L'expérience dure environ 20 heures et correspond à 1012 points de mesures des paramètres biochimiques. On considère le début du bio-procédé comme étant t=0 heure (0h). On cherche à détecter trois états physiologiques principaux : (1) la fermentation (production d'éthanol), (2) la diauxie (production d'acide) et (3) l'oxydation (production de biomasse).
Chaque état physiologique est composé d'une ou plusieurs classes. Il y a 22 paramètres biochimiques et chacun d'eux a donc 1012 éléments. Pour chaque élément, chaque paramètre est testé pour voir s'il est en conflit ou non avec la majorité des autres

paramètres. Le conflit est calculé avec la mesure présentée dans le paragraphe 4.3. Expérimentalement, nous avons pris un seuil de conflit égal à 0,3. Ainsi si un paramètre P1 a un conflit avec un paramètre P2 supérieur à 0,3 alors P1 est en conflit avec P2. Si P1 est en conflit avec la majorité des autres paramètres alors P1 est considéré comme non pertinent. A titre d'information si l'on voulait utiliser la mesure de conflit classique il faudrait prendre un seuil de 0,9, autrement quasiment tous les paramètres seraient en conflit les uns avec les autres. Pour chaque échantillon, nous combinons les paramètres pertinents. Pour la classification, nous avons choisi le maximum de plausibilité pour avoir un comportement optimiste. Pour le calcul des masses d'évidence, nous avons normalisé les degrés d'appartenance fournis par la classification LAMDA. Il faut noter que nous avons testé une autre méthode pour calculer les masses d'évidence à partir de ces degrés d'appartenance. Cette méthode développée par Desachy et al. [Desachy *et al.*, 1996] permet en effet d'affecter une masse d'évidence à l'union de toutes les classes; cette masse d'évidence affectée à l'union de toutes les classes représente l'ignorance de la source d'information. Cependant dans cette application, les calculs effectués par cette méthode montrent que les masses d'évidences affectées aux classes sont très faibles (de l'ordre de 10^{-2} ou 10^{-3}) comparativement à la masse d'évidence de l'union des classes (de l'odre de 10^{-1}) et ce, quelque soit le paramètre biochimique. Ainsi avec cette méthode, tous les paramètres présentent un degré d'ignorance important. Nous avons donc préféré utiliser la "simple" normalisation des degrés d'appartenance pour tester la pertinence des paramètres. Les premiers résultats ont été présentés dans [Régis *et al.*, 2004a].
Les résultats sont particulièrement intéressants. Premièrement, concernant la classification en elle-même, on constate d'une part, l'apparition de nouvelles classes auparavant absentes à certains moments de l'expérience, et d'autre part la disparition d'autres classes autrefois présentes à d'autres moments de l'expérience. La disparition de classes est due à l'élimination de certains paramètres considérés comme non pertinents et qui influençaient fortement la classification, tandis que l'apparition de certaines classes peut correspondre à des sous-états physiologiques.
Deuxièment, la pertinence des paramètres (pour plus de précision sur ces paramètres, voir l'annexe 1) permet de fournir une première analyse :

- 8 paramètres sont considérés comme non pertinents à la fin de l'expérience (entre 18h et 20h). On peut citer par exemple les paramètres suivants : ajout de base, tension électrique, luminance, rO_2 et pH . Ce manque de cohérence est confirmé par les experts. En effet, ce phénomène est bien connu par ceux-ci et provient de la décroissance et de la mort des micro-organismes. Cette mort entraîne une situation chaotique qui explique le nombre élevé de paramètres non pertinents.
- certains paramètres sont considérés comme non pertinents au milieu de l'expérience (QR ou Quotient Respiratoire, O_2 mesuré, action chaude). Or lorsque l'on analyse les signaux, on contaste que ce manque de pertinence correspond exactement à l'apparition de pics ou de singularités qui semblent souvent inexplicables et incohérents pour les experts. En d'autres termes, ces singularités sont des artefacts. Ainsi cette méthode est capable de détecter ces artefacts et de les éliminer de la classification. Par exemple, sur la figure 1, pour le paramètre QR la singularité localisée à 5,35h est un artefact et n'est pas pris en compte par la

classification. De même sur la figure 1, pour le paramètre O_2 mesuré (oxygène mesuré) l'artefact qui apparait à 15,3h est éliminé de la classication. La classification tend donc à détecter les fautes, agissant comme un système intelligent. Cependant toutes les singularités ne sont pas systématiquement éliminées : en effet un paramètre multifractal comme le CO_2 mesuré (dioxyde de carbone mesuré, voir figure 2) est considéré comme un paramètre pertinent pendant toute la durée de l'expérience bien qu'il possède un grand nombre de singularités. La méthode tend donc à caractériser les singularités en fonction des propriétés analytiques des paramètres pour éliminer les vrais artefacts.

Par ailleurs, si l'on effectue une ACP sur cette expérience, il est difficile de tirer une conclusion quant à la non pertinence des paramètres (voir figure 3). On constate que les paramètres QR et CO_2 mesuré (qui est nommé "CO2_mes" sur la figure 3) sont plutôt corrélés et pourtant le CO_2 mesuré est toujours pertinent alors que le QR ne l'est pas pour un point localisé. Ceci est dû au fait que l'ACP n'est pas réalisable sur un point mais sur un ensemble de points. D'autre part une ACP sur des intervalles de temps pourrait être plus significative mais le problème de la définition des bornes de ces intervalles reste entier. En effet un désaccord ne coïncide pas forcément avec une singularité. Ainsi le paramètre action chaude est en conflit avec la majorité des autres paramètres et n'est donc pas pertinent au début de l'expérience d'après la méthode présentée dans ce papier (voir figure 2). Cette observation est confirmée par les microbiologistes qui considérent cet intervalle de temps comme une période de calibrage de certains paramètres (comme l'action chaude) où un tâtonnement est souvent nécessaire. Or si l'on avait voulu déterminer un intervalle pour l'ACP, on aurait des difficultés à trouver ses bornes car au début de l'expérience, le paramètre action chaude ne présente aucune singularité. Et même si le début et la fin d'un désaccord correspondent à des singularités, la décorrélation d'un paramètre avec les autres n'implique pas nécessairement qu'il soit en conflit avec ceux-ci. Ceci étant, le fait que l'action chaude (nommée "Act∘_chd" sur la figure 3) soit décorrélée par rapport à certain signaux (le pH, la tension électrique qui est nommée "U_(volt)ppe" sur la figure 3, etc) et anticorrélée par rapport à d'autres (le QR, le CO_2 mesuré) peut traduire la tendance au conflit au début de l'expérien ce. Cependant, les résultats de l'ACP sont plus difficiles à interpréter car, d'une part, ils ne montrent pas que le conflit de l'action chaude avec les autres paramètres disparait dans la suite de l'expérience, et d'autre part, l'action chaude reste trés corrélée à des signaux comme l'agitation ou la température, qui eux, ne sont pas du tout en conflit avec le reste des paramètres au début de l'expérience. Ainsi les méthodes statistiques comme l'ACP fournissent des résultats qui semblent plus difficiles à interpréter pour l'évalutation du conflit entre paramètres et pour la localisation temporelle de ces conflits. Ainsi la méthode de classification de Dempster-Shafer utilisant la pertinence des paramètres présente plusieurs avantages: elle confirme les connaissances des experts concernant la situation chaotique à la fin de l'expérience, et ce, de façon totalement indépendante de ces connaissances; et elle agit comme un système intelligent en détectant et en éliminant les artefacts.

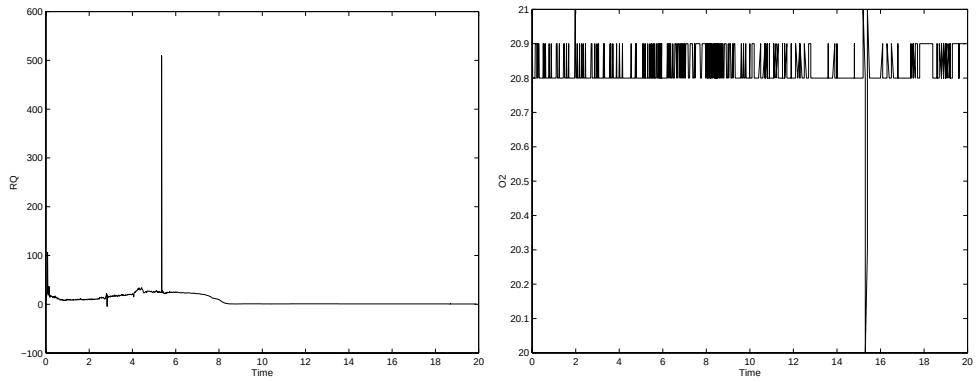

FIG. 1 – *Un artefact apparaît respectivement dans le QR à t=5,35h (image de gauche) et dans l'O_2 à 15,3h (image de droite), mais il sont tous les deux éliminés de la classification.*

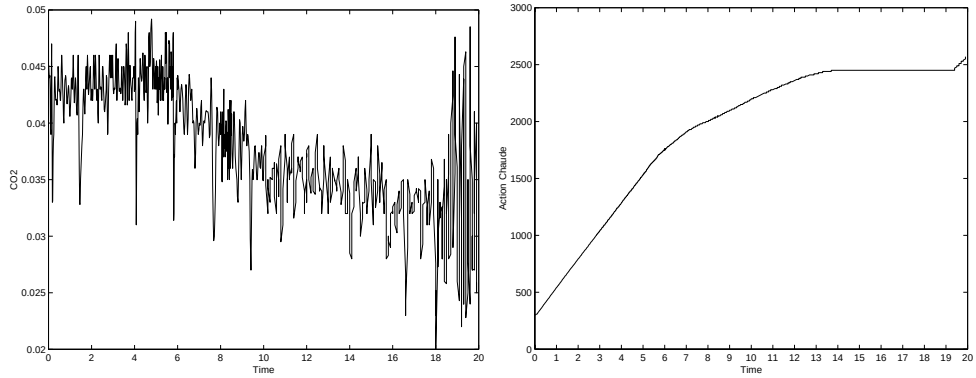

FIG. 2 – *Le CO_2 mesuré (image de gauche) est multifractal mais est toujours considéré comme pertinent. L'action chaude (image de droite) n'est pas pertinente de 0 à 1h mais ne présente aucune singularité dans cet intervalle de temps.*

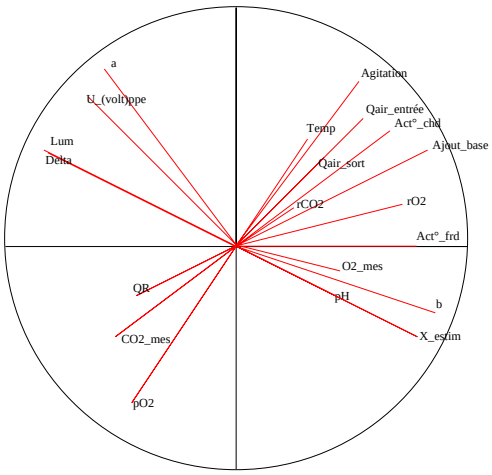

FIG. 3 – *ACP de l'expérience*

6 Conclusion

Dans ce papier nous avons présenté une méthode de classification basée sur la théorie de Dempster-Shafer et sur LAMDA pour l'analyse de paramètres biochimiques. Les paramètres biochimiques de la fermentation sont caractérisés par leur pertinence. La pertinence d'un paramètre est basée sur la notion de conflit. Une autre mesure du conflit a été proposée et fournit de meilleurs résultats que la mesure classique du conflit. Les premiers résultats montrent que l'évaluation de la pertinence par cette méthode confirme les analyses microbiologiques et qu'elle agit comme un système expert concernant les artefacts. Les futurs travaux concerneront l'analyse de la pertinence sur des intervalles de temps (et non simplement une analyse point par point) et l'utilisation de cette méthode à d'autres applications utilisant plusieurs sources d'information. Par ailleurs, des tests supplémentaires utilisant d'autres méthodes de calcul des masses d'évidences doivent êtres effectués et analysés. Nous remercions l'équipe fermentation du LBB INSA pour avoir mis à notre disposition les données expérimentales et pour leur aide à la validation des résultats.

Références

[Aguilar-Martin *et al.*, 1980] J. Aguilar-Martin, M. Balssa, et R.-L. De Mantras. Estimation récursive d'une partition. exemple d'apprentissage et auto apprentissage dans **R**. Technical Report 880139, LAAS-CNRS, 1980.

[Aguilar-Martin *et al.*, 1999] J. Aguilar-Martin, J. Waissman-Vilanova, R. Sarrate-Estruch, et B. Dahou. Knowledge based measurement fusion in bio-reactors. In *IEEE EMTECH*, Mai 1999.

[Chan *et al.*, 1989] M. Chan, J. Aguilar-Martin, N. Piera-Carreté, P. Celsis, et J.-P. Marc Vergnes. Classification techniques for feature extraction in low resolution

tomographic evolutives images: application to cerebral blood flow estimation. In *12th Conf. GRESTI*, 1989.

[Dempster, 1968] A. P. Dempster. A generalisation of bayesian inference. *Journal of the Royal Statistical Society*, 30:205–247, 1968.

[Desachy *et al.*, 1996] J. Desachy, L. Roux, et E.-H. Zahzah. Numeric and symbolic data fusion: a soft computing approach to remote sensing images analysis. *Pattern Recognition Letters*, 17:1361–1378, 1996.

[Dubois et Prade, 1992] D. Dubois et H. Prade. On the relevance of non-standard theories of uncertainty in modeling and pooling expert opinions. *Reliability Engineering and System Safety*, 36(2), 1992.

[Dubois et Prade, 2004] D. Dubois et H. Prade. On the use of aggregation operations in information fusion process. *Fuzzy Sets and Systems*, 142:143–161, 2004.

[Nakkabi *et al.*, 2002] Y. Nakkabi, S. Regis, J. Desachy, A. Doncescu, et G. Roux. Apport de la transformée en ondelettes pour affiner les résultats de classifications. In *IXe Rencontre de la Société Francophone de Classification*, pages 287–291, Toulouse, Septembre 2002.

[Piera-Carreté *et al.*, 1990] N. Piera-Carreté, N.-P. Desroches, et J. Aguilar-Martin. Variation points in pattern recognition. *Pattern Recognition Letters*, 11:519–524, 1990.

[Piera-Carreté et Aguilar-Martin, 1991] N. Piera-Carreté et J. Aguilar-Martin. Controling selectivity in non-standard pattern recognition algorithm. *Trans. in Syst. Man and Cybernetics*, 21:71–82, 1991.

[Regis *et al.*, 2003] S. Regis, J. Desachy, A. Doncescu, et J. Aguilar-Martin. Comparaison de classifications non supervisées de données biotechnologiques. In *Xe Rencontre de la Société Francophone de Classification*, Neuchâtel, Suisse, Septembre 2003.

[Régis *et al.*, 2004a] S. Régis, J. Desachy, et A. Doncescu. Evaluation of biochemical sources pertinence in classification of cell's physiological states by evidence theory. In *FUZZ-IEEE*, Budapest, Juillet 2004. accepté.

[Regis *et al.*, 2004b] S. Regis, L. Faure, A. Doncescu, J.-L. Uribelarrea, L. Manyri, et J. Aguilar-Martin. Adaptive physiological states classification in fed-batch fermentation process. In *9th IFAC Intern. Symp. on Computer Application in Biology*, Nancy France, Mars 2004.

[Shafer, 1976] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, New Jersey, 1976.

[Smets, 1990] P. Smets. The combination of evidence in the transferable belief model. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, (12):447–458, 1990.

[Steyer, 1991] J.-P. Steyer. *Sur une approche qualitative des systèmes physiques : aide en temps réel à la conduite des procédés fermentaires*. PhD thesis, Université Paul Sabatier, Toulouse France, Décembre 1991.

[Waissman-Vilanova *et al.*, 1998] J. Waissman-Vilanova, J. Aguilar-Martin, B. Dahhou, et G. Roux. Généralisation de degré d'adéquation marginale dans la méthode de classification lamda. In *VIe Rencontres de la Société Francophone de Classification*, 1998.

[Waissman-Vilanova *et al.*, 2000] J. Waissman-Vilanova, R. Sarrate-Estruch, B. Dahou, et J. Aguilar-Martin. Wasterwater treatment process supervision by means of fuzzy automaton model. In *IEEE ISIC*, Juillet 2000.

[Waissman-Vilanova, 2000] J. Waissman-Vilanova. *Construction d'un modèle comportemental pour la supervision de procédés: application á une station de traitement des eaux*. PhD thesis, LASS - CNRS, Novembre 2000.

[Yager et Rybalov, 1998] R. Yager et A. Rybalov. Full reinforcement operators in aggregation techniques. *IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics*, 28(6), 1998.

[Zadeh, 1984] L. A. Zadeh. Book review: A mathematical theory of evidence. *AI Magazine*, 5(3):81–83, 1984.

Annexe 1

Dans cette annexe, nous donnons quelques explications pour certains des paramètres présentés.
L'action chaude permet de réguler la température du milieu. L'agitation correspond à la vitesse d'un rotor qui permet de garder l'homogénéité du milieu. L'ajout de base permet de réguler le pH. Le rCO_2 et le rO_2 correspondent respectivement à la vitesse de production de dioxyde de carbone, et à la vitesse de consommation de l'oxygène. Le CO_2 mesuré et l'O_2 mesuré correspondent à la mesure directe de dioxyde de carbone et d'oxygène présents dans le milieu. La luminance dépend de l'intensité lumineuse du milieu et mesure la concentration de la biomasse (la biomasse est la quantité de micro-organismes) dans le milieu. Le QR est le rapport du rCO_2 sur le rO_2. La tension électrique mesure l'activité électrochimique des micro-organismes.

Summary

For analysis and modelling of the biotechnological process we must look deeper to the biological systems. The cell metabolism and the resulting kinetic form a complex process which could not be modelling completely by a non linear differential system. The goal of all modelling is either the biocontrol or finding the physiological states. We want to detect the physiological states using a small number of measured signals. We present in this paper the analyze of biochemical parameters using the evidence theory. The evidence theory is also used to characterize the pertinence of the parameters. This pertinence is based on the notion of conflict. We show that our measure of conflict based on a distance provides more coherent results as the classical distance in somes cases, and particularly in the microbiological process which is presented.