Bioprocess Diagnosis based on the empirical use of distance measures in the theory of belief functions

Sébastien REGIS[a], Andrei DONCESCU[b], Nathalie GOMA[c]

[a]*LAMIA*
French West Indies and Guyana University
97159 Pointe--Pitre, Guadeloupe France
Email: sregis@univ-ag.fr, http://lamia.univ-ag.fr
[b]*LAAS-CNRS 7*
Avenue du Colonel Roche 31077 Toulouse Cedex 4
Email: adoncesc@laas.fr, http://www.laas.fr
[c]*IPBS-CNRS*
Route du Narbone Toulouse Cedex 4, France

Abstract

Microorganisms play a central role in the production of a wide range of industrial chemicals, enzymes and antibiotics. The rate of product formation in a given industrial process, is directly related to the rate of biomass formation which is influenced directly or indirectly by a whole host of different environmental factors. In this article we propose to use distance measures between basic belief assignment in the context of the belief functions theory, in order to diagnosis the relevance of bioprocess sensors and actors which measure the environmental factors.

Keywords: Theory of belief functions, conflict, distance measures, relevance of source, diagnosis, bioprocess.

1. Introduction

The establishment of reliable processes with increased efficiency and cost reduction is of primary importance in the fermentation industry. Nowadays, the increase of the number of sensors in biotechnology field provides an important amount of heterogenous data. This data with various natures and

various levels of precision require the use of automatic method of data fusion. The goal of data fusion is to diagnosis here the correctness of the bioprocess. The role of fermentation diagnosis is to detect the anomalies able to modify the optimal conditions maximizing out of product, in our case biomass. Anomalies can occur in biochemical parameters such as temperature, aeration, pH and dissolved oxygen. Monitoring and diagnosis of bioprocesses has been tackled in three ways:

1. Adaptive state estimators: These approaches are designed to adapt to the time varying characteristics of the process, for example to the changes in growth and metabolite expression rates when the nutrient levels are depleted. The measurements are used in conjunction with the process model and need to be carefully tuned to achieve accurate reconstruction of fermentation parameters and states. There are two mains reasons. First, because the dynamics of biological processes are both non-linear and non-stationary; secondly, because classical methods have proved inadequate in describing the overall behavior of biological process.
2. Artificial intelligence based algorithms: The second approach to the development of monitoring and fault diagnosis strategies is based on artificial intelligence based algorithms. Such approaches rely on the construction of a qualitative and quantitative database of regular and faulty modes of plant operation. The event-tracking is classified as "normal" or "faulty" using a supervised classifier. Heuristics based ex-

pert systems rely on capturing knowledge and know-how) on the growth conditions of microorganism, and take into account the different fault occurrences and process variable interactions that the plant personnel can envision, by capturing them a rule base. [1][2].

3. Model-based Statistical Signal Processing : In these approaches, statistical models are developed using event-tracking data collected during the routine normal operation of fermentation. The data from the batch is compared with the template of normal conditions established in the statistical model and diagnosed for process upsets and sensor failures.

The contradictions between observation and predictions allow more than the task detection but provide information on the localization of these faults. Basically, the predictions obtained from models (analytical or logical) lead to contradictions with the observations produced by the different sources of information if these are not correct. This reasoning by absurd lets us regroup these conflicts in order to localize the faults. The detection of conflicts is the first step in the diagnostic. The second consists in generating the diagnosis from the set of conflicts.

In this frame work the theory of belief functions used in this paper is an Artificial Intelligence Method of Diagnosis. The originality of the approach presented in this paper is its capacity to manage imprecise and uncertain information. The final purpose of our approach is, first, to have an automatic method to evaluate the relevance of an information source, instead of using the subjective knowledge of an human expert, and secondly, to improve

the result of classification (by detecting non-relevant sources of information). More precisely, we use the notion and values of distance measures in order to assess the relevance of a source. This approach enables to estimate the relevance without *a priori* informations on the source: the estimation of the relevance is based only on the intrinsic informations. We use two measures of conflit: a measure based on a norm between the mass function of the sources [3, 4] and a measure called the Jousselme distance [5]. We test the evaluation of the relevance with the two measures on a practical application of fermentation bioprocess and the results show that this approach improves the results of classification.

The paper is organized as follows: section 2 presents the context and problematic of bioprocess diagnosis, section 3 presents the notion of relevance; section 4 provides the basic notion of the theory of belief functions. In the section 5 we make the connection between the relevance and the theory of belief functions and section 6 presents the first experimental results for a bioprocess.

2. Bioprocess Diagnosis

There are several kind of bioprocesses. One of the most well known is the batch, but there are also fed-batch and single CSTR (chemostat) bioreactors which also fed with sterile nutrient medium. The difference between them is in the manner of operate without or with external sources of biomass after inoculation. When we speak of by external source of biomass, the biomass

introduced into the bioreactor after the inoculation. After this process of inoculation, the culture is maintained at conditions that are compatible with growth (e.g. at suitable temperatures) and often kept in an agitated state.

Three biological reactions need to be diagnosed in view of control :

Growth :

$$Glucose + Salts + Vitamins + O_2 \rightarrow Y_{S,X} Biomass + CO_2 + H_2O \quad (1)$$

Ethanol Production :

$$Glucose \rightarrow Y_{S,P} Ethanol + CO_2 \quad (2)$$

Maintenance :

$$Glucose + O_2 \rightarrow CO_2 + H_2O \quad (3)$$

where Y represents the yield.

Depending on the organism being cultivated, the fermentation is typically carried out at volume ranging from a few liters up to a few hundred thousand liters, and lasting for a period of several hours or up to a few weeks. Furthermore, there may be additional requirements for amino acids, vitamins, purines, or pyrimidine. It may be necessary to aded precursors according to the metabolites that we want to obtain. Oxygen may or may not be required as the terminal electron acceptor, and the fermentation will need to be carried out at an appropriate pH. The medium components used to satisfy an

organism's nutritional requirements are partially influenced by the nature of the fermentation process used developed.

The biological system studied in this article is the yeast *Saccharomyces Cerevisiae*. Yeast is one of the smallest eukaryotic systems sequenced and is unparalleled for the level of molecular investigations that have been carried out and the range of possible manipulations. It is an ideal target for a comprehensive study at the system-level. Two directions have been explored:

1. *on-line* analysis : it does not permit diagnosis in an instantaneous manner or with certainty regarding the physiological state of the yeast.
2. *off-line* analysis : makes it possible to soundly characterize the current state, but unfortunately after too late to take into account this information or to adjust the process on the fly by actions of regulators allowing to adjust some critical parameters such that pH, temperature (addition of basis, heats, cooling).

To remedy these drawbacks, computer scientists in collaboration with micro-biologists have developed tools for supervised control of the bioprocess. They use the totality of information provided by the sensors during a set of sample processes to infer some general rules to which the biological process obeys. These rules can be used to control the next processes. This is exactly the problem we tackle in this paper. To sum up, our application focuses on the evolutive behavior of a *bio-reactor* (yeast fermentation), that is to say an evolutive biological system whose interaction with the physical world, as described by pH, pressure, temperature, etc..., generates an observable

reaction. This reaction is studied by way of a set of sensors providing a large amount of (generally) numerical data.

Physiological state diagnosis uses the relevant information provided by sensors in a process of classification. This "relevant classification" associates classes to physiological states. Faults or abnormal behavior means apparitions of physiological states different from the fermentation goals. Therefore our goal is to avoid the reinforcement of the faults or abnormal behavior in the classification process.

The data used in this article was obtained from the biomass production. *Saccharomyces Cerevisiae* was studied under oxidative regime (i.e. no ethanol production). Two different protocols have been applied: a batch procedure that is followed by a continuous procedure. The batch procedure is composed by a sequence of biological stages. This phase can be thought of a start-up procedure. Biotechnologists state that the behavior in the batch procedure induced influences later phenomena in the continuous phase. So complete knowledge of the batch phase is of great importance for the biotechnologist. The traditional way of getting acquainted with such knowledge is at present carried out through offline measurements and analysis which most of the time produce results when the batch procedure has ended, thus lacking real time performance. Contrast the proposed methodology allows for real time implementation. This example deals with the batch procedure. The expert chooses among the set of available on-line signals which, according to the expert knowledge contain the most relevant information to diagnose the

physiological state:

1. DOT : partial oxygen pressure in the medium.
2. O2 : oxygen percent in the output gas
3. CO2 : carbon dioxide percent in the output gas
4. pH.
5. OH- ion consumption : derived from control action of the pH regulator and the index of reflectivity.

The consumption of negative OH ions is evaluated from the control signal of the pH regulator. The actuator is a pump, switch on by an hysteresis relay, that inoculates a basic solution (NaOH). The reflectivity, which is measured by the luminance, seems to follow the biomass density. Nevertheless its calibration is not constant and depends on the run.

The analysis of signals provided by sensors and actors allows to build up a system of physiological state diagnosis. In practice, we have to make a classification of the biochemical parameters: the aim is for a class (or a group of classes) obtained from classification to correspond to a physiological state. The detection of the physiological state makes it possible to control and optimize the bioprocess. The classification involves segmenting the biochemical parameters (which are time series) so that a class or a group of consecutive classes correspond to a precise physiological state. In fact, classification can be made "manually" by an expert in microbiology (see figure 1). Our goal is to automize to the analyze given by the expert (see for example [6, 7, 8]).

As we said above, the microbiologist experts need to better control the physiological states of the micro-organisms but also the biochemical parameters which enable the characterization of the physiological states. Indeed, the microbiologists have knowledge no or a little about the pertinence of the biochemical parameters, and this knowledge is mainly *empirical* and *subjective*. Moreover this knowledge tends to deal with no more than five parameters at a time while there are generally at least 12 parameters. A lot of information is then either redundant, unexploited or even erroneous. So it is important to find tools to analyze these parameters in order to confirm or cancel the expert knowledge and then add or remove biochemical parameters from the classification.

In fact, the approach proposed in this article tries to answer this question: *Is it necessary to use all the biochemical parameters to make a diagnosis ?* The answer is not obvious, as one [9] shows that less of the half of the available parameters are actually used in some applications. And even if the expert in microbiology should answer yes to the question, we have seen that they only use some of these signals. That's why it's interesting to have an automatic method to evaluate the relevance of the parameters. This automatic method is based on the theory of belief functions and particularly on the notion of distance measures between belief functions distributions.

Figure 1: Physiological states provided by the expert for the bioprocess. x-axis is the time in hours, y-axis represents the values of the parameters (values have been normalized). 4 parameters were used: pH (measure of the acidity), rCO2 (speed production of carbon dioxide), rO2 (speed consumption of oxygen), and the luminance (which traduces the biomass production). There are 3 states: the state 1 (the fermentation), the state 2 (the diauxy) and the state 3 (the oxidation)

3. Diagnosis using Artificial Intelligence-Based System

The Diagnosis approach based on Artificial Intelligence supposes the available knowledge is limited to a set of observations (possibly uncertain or imprecise) of some variables and the problem of diagnosis is reduced to a problem of decision. The adaptation of a decision system to non-steady states is a fundamental problem in diagnosis but not directly linked to a system of representation of the uncertainty. In this article we characterize the partial knowledge from some observations which are imprecise and uncertain. Let us present the notion of relevance of data and some approaches used to manage it.

3.1. Notion of relevance

The notion of relevance (or pertinence) remains very imprecise as there is no single definition in the fields of computer science. However, this notion of relevance (but also and especially the notion of irrelevance) is widely used in the fields of computer sciences: classification [10][11],fault detection [12], intelligent websearch engines [13][14], and so on. The definition of relevance depends on its use and the framework. For Lazo-Cortes and Ruiz-

Schulcloper [10], the notion of relevance is linked to the ability to distinguish several classes. In the fields of information science, for Spoerri [15], the notion of relevance of a source depends on the number of information retrieval systems which find this source. For Zadeh [13], the global relevance of a source depends on the information given by the other sources. For Paltoglou et al. [14], the relevance of a source is based on the relevance of the data (i.e. documents) of this source. Blum and Langley[11] propose at least 5 different definitions of the relevance and Pichon et al [16] modelize the relevance (and also the truthfulness) by using a mass function in the framework of the theory of belief functions.

Relevance can be defined generally as the relation to the matter at hand, the nature of which is linked to the objective. This definition is based on the encyclopedic definition [17][18]. We suggest a first definition of relevance.

Definition 1

A source is relevant if:

- it does not provide aberrant results
- it provides meaningful information for the classification
- it implies a decision which agrees with most of the other sources

The first requirement shows that the source doesn't produce artifacts. The second one is linked to the fact that the objective is to find a correct classification. The last is based on the assumption that the majority of the sources

reflects the truth at least to some extent: this last one is linked to the application. For example in the case of a bioprocesses (we will present the experimental results in these fields) we have assumed that on the majority of the biochemical sources, because all the sources should reflect the same biological situation.

3.2. Methods of evaluation of relevance

Several automatic methods exist to evaluate the relevance of a source. For instance Blum and Langley [11] propose classifying them into 4 groups:

- embedded approaches. These are methods with recursive classifications where the evaluation of relevance is made in the same time as the classification. Methods using logical rules [11] for instance, belong to this class.

- filter approaches. This is is note related to signal processing. The notion of filtering comes from the fact that these approaches use mathematical tools, independently of classification. Principal Component Analysis (PCA) for example can be viewed as a filtering approach to select relevant features.

- wrapper approaches. These are based on the analysis of the classification and thus they are made after the classification. The method proposed by Dubois and Prade [19] and as well as our approach belong to this category of approach.

- weighting methods. These are methods using weighting functions to provide a degree of relevance for each source. Neural networks can be used for this approach.

Other approaches such as *Branch & Bound* [20], voting [19] or the Felix's method [21] can be viewed as methods which either implicitly or explicitly evaluate the relevance. We must note that most of these methods give a global evaluation of each source and not a local evaluation. The notion of dynamical relevance is more adapted to bioprocesses as the relevance of biochemical parameters can vary according to the evolution of the bioprocess. Thus we propose to use the belief functions theory to evaluate the relevance of biochemical parameters.

4. Theory of belief functions

4.1. Basic notions

Belief functions theory can be viewed as a generalization of the Bayesian theory which take into consideration uncertainty and partial knowledge. It was introduced by Dempster [22] and was mathematically formalized by Shafer [23].

We consider all hypotheses; this set is called the *frame of discernment* and is denoted Θ. All the elements of the frame are exclusive and are called *singletons*. The belief functions theory works on the set of subsets (or assertions) A of Θ. This set of subsets of Θ is denoted 2^{Θ}. A can be a union of several

singletons or one singleton. A mass function $m(.)$ is then defined from 2^Θ to [0,1] with the following properties:

$$\sum_{A \subseteq \Theta} m(A) = 1 \qquad (4)$$
$$m(\emptyset) = 0$$

The belief and plausibility functions are defined from 2^Θ to [0,1]:

$$Bel(A) = \sum_{B \subseteq A} m(B) \qquad (5)$$
$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B)$$

The Dempster rule allows to merge the mass function of two sources (or more):

$$(m_1 \oplus m_2)(A) = \frac{1}{1-K} \sum_{B \cap C = A} m_1(B) \cdot m_2(C) \qquad (6)$$
$$A, B, C \in 2^\Theta$$

K (called Dempster's conflict) is defined as:

$$K = \sum_{B \cap C = \emptyset} m_1(B) \cdot m_2(C) \qquad (7)$$

The denominator $(1-K)$ is here to enable the normalization. K is a measure of a kind of conflict between the sources. If K is equal to 1, the fusion is

undefined. For several sources, it is possible to compute K iteratively using the conflict between two sources then between these two sources and a third source etc.

4.2. About the conflict in the theory of belief functions

Various studies have been done on the notion of conflict defined by the theory of belief functions. Several alternatives merging rules [24][25][26][27][28] have been proposed to overcome the pre-supposed erroneous results generated by the Dempster's conflict used in the combination from theory of belief functions (see for instance the famous example proposed by Zadeh in [29]). However it has been show [30] that this counter-intuitive example proposed by Zadeh, showing that Dempster's conflict K can lead to erroneous conclusion, does not come from erroneous properties of the Dempster rule of combination themselves, but rather from an erroneous use of the theory of belief functions. Moreover, erroneous results can come from the combination of subproblems that ought to be handled independently [31][32]. On the other hand, the fact that the Dempster's conflict of two identical belief functions is not null has been studied and explained [33][34]. Furthermore, the notion of conflict itself K has been reevaluated [33][35][36]: the Dempster's conflict K is not really a conflict measure between the basic belief assignments and can be interpreted qualitatively as an indicator of compatibility between two hypothesis. Nowadays, after the proposition of new formulas for

the combination in belief functions theory, one of the trends is to propose and classify several kinds of distance measures between bodies of evidence[37][38]. A possible distinction between existing measure distance is based on what kind of distances are measured by these measures. There are at least two possibilities [39]: some measures distance measure the degree to which two bodies are different (as the Minkowski family of distance belief functions[38], based on metrics) and other distances (as Dempster's conflict) measure the degree to which they are compatible. Different measures measure different types of distance [39][38]. The choice of a distance measure should be based on the application and the fusion's aim.

5. Diagnostic of relevance and the theory of belief functions

5.1. Use of the theory of belief functions for relevance

We propose to use distance measures from belief functions theory to evaluate the relevance of the biochemical parameters (which are here the sources of information). One of the first to introduce the conflict of belief functions theory to compare sources of information was Schubert [31, 40]. He proposed using the Dempster's conflict between sources in order to regroup the sources which have small differences and thus make clusters of sources of information. The aim of theses clusters is to have a more coherent fusion of sources. Moreover, in belief functions theory it is important [41] to select the sources before making the fusion . Régis et al. [3, 4] then Chebbah et al. [42] and Martin et al. [37] have independently proposed using distance measure to

evaluate the quality of the sources of information. Régis et al. call this quality relevance whereas Chebbah et al. and Martin et al. call it reliability, but the two notions are quite similar. By computing the distance measure two by two between the sources of information, it is possible to see which sources are in consensus and which sources are not. For a given source, the mean of the distance measure (two by two with the other sources) is computed and if this mean is lower than a threshold $\tau \in [0, 1]$ (a priori, $\tau = 0.5$ but empirically its value can vary) it is considered as relevant, otherwise it is considered as non pertinent. Thus as we said above, we have made the assumption that the majority of the sources are valid providing that all the sources observe the same situation. It is then possible to characterize with a certain flexibility the pertinence of the sources. This characterization is made for each sample, that is to say in a local way. This local characterization is more significant than a global characterization which does not take into account the evolution of the experience over time. Thus if a source is estimated as non-relevant, it is weakened [23] before the fusion of the sources. The approach can be summarized as follows:

- For each given time t:
    1. Characterization of the relevance
        - For each parameter P
            - calculating of the distance measure two by two between the parameter P and the other parameters

- if the mean of the distance measures P is lower than a threshold τ,it is relevant
- otherwise it is not relevant

2. Fusion of the information by the Dempster combination of all the parameters with:
   - for all the relevant parameters all the masses remain the same
   - for the irrelevant parameters, new masses are calculated and updated with a weight of 0.5:

$$m(A) \leftarrow 0.5 \times m(A) \quad \forall A \subset \Theta \tag{8}$$

$$m(\Theta) \leftarrow 1 - 0.5 \times (1 - m(\Theta)) \tag{9}$$

We propose to use a *binary* approach (relevant/not relevant) for the data source by using a reducing weight arbitrarily equal to 0.5 but one can note that there are approaches to evaluate and optimize the *degree* of quality (or relevance/relaibility) for a data source [37]. Another approach which can be used is the discounting of the data source until it enables to have results *fitting* to a predefined level of conflict; more details can be found in the work of Schubert [39].

5.2. Use of distance measures

As we have said above, the choice of a distance measure depends on the application and the aim of the data fusion. Our approach is based mainly

on the definition of the relevance (see definition 1 in 3.1) particularly for the third point:"a source is relevant if it implies a decision which agrees with most of the other sources". This point clearly implies a notion of similarity between data sources and led us to choose measures distances which let us to measure the degree to which two bodies are different. The Dempster's conflict is not well adapted for this application because it does not measure similarity or difference between two belief function distributions. We propose to use two distance measures, both from the Minkowski family of distance belief functions[38].

The first is based on the usual norm called norm 1 [3, 4](see also [43]) and is defined as follows:

$$conf_1(S_i, S_j) = d_{met}(S_i, S_j) = \frac{1}{2} \sum_i |m_1(A_i) - m_2(A_i)| \qquad (10)$$
$$A_i \in 2^\Theta$$

where $m1$ and $m2$ are the mass functions for the information sources S_i et S_j. The factor $\frac{1}{2}$ is a factor of normalization.

This distance measure is easy to use, but its main drawback is that it can be used only if the mass function of each of the sources are all distributed on the same focal elements (in practice it is often the case, but one can find actual or simulated cases where the mass functions of the sources do not work on the same elements).

The second distance measure was used by Chebbah et al. [42] and is called

the Jousselme distance [5]. This enables us to reflect on the specific functions of belief since this distance used the Jaccard coefficient $\frac{|A_i \cap A_j|}{|A_i \cup A_j|}$ where A_i and A_j are two focal elements. This Jaccard coefficient takes into account the cardinality of focal elements. A matrix of the Jaccard coefficient D is well defined on the set 2^{Theta}, which makes this specific distance belief functions. Jousselme distance is given by:

$$d_{jous}(m_1, m_2) = \sqrt{\frac{1}{2} \cdot (m_1 - m_2)^t D(m_1 - m_2)} \tag{11}$$

with

$$\begin{aligned} D(A_i, A_j) = 1 \quad if \quad A_i = A_j = \emptyset \\ = \frac{|A_i \cap A_j|}{|A_i \cup A_j|} \quad \forall A_i, A_j \in 2^{\Theta} \end{aligned} \tag{12}$$

where m_1 and m_2 are two mass functions of the sources S_1 et S_2. The distance measure between S_1 and S_2 is given by:

$$conf_2(S_1, S_2) = d_{jous}(m_1, m_2) \tag{13}$$

The obvious advantage of this measure is that it can be used without condition on the distribution of mass functions on the focal elements. It may therefore be used in all real cases where the sources are working or not on the same focal elements. The value of distance Jouselme will be preferred to the distance metric for experimental tests.

So whatever the distance measure used, the average distance of a source S_i (from other sources, more precisely from the belief function distributions of the other sources) is defined as follows:

$$ConfMean(S_i) = \frac{1}{n-1}.\sum_{j,j\neq i} conf(S_i, S_j) \tag{14}$$

where n is the number of sources.

6. Experimentation

6.1. Material: bath fermentation bioprocess

The analyzed experiment is a batch bioprocess (using the yeast Saccharomyces Cerevisiae) which takes about 20 hours and corresponds to 1012 points of measurement of the biochemical parameters. We consider the start of the bioprocess as t = 0 hours (0h). Recall that we are trying to identify three main physiological states (see figure 1):

1. State 1: fermentation (ethanol). It goes up about for 9h (from 0h to 9h) which represents a total of 590 points measured.
2. State 2: dioxidation . This state starts at about 9h and ends at 9h46 which is about 33 points. This is the smallest physiological state (time and amount of data) and it is most difficult one to characterize among the three states.
3. State 3: oxidation (biomass). It begins at 9h46 and ends at 20h, and represents 389 points.

There are 22 biochemical parameters and each of them has 1012 elements. So there is time series with 1012 points for each of them.

6.2. Methods: mass functions for theory of belief functions

Preliminary works on this kind of bioprocess can be found in [3][4][44]. We use two different kinds of computation for the mass function, one for each kind of bioprocess. In order to compute the mass function, we use the method proposed by Denoeux [45] because in this bioprocess it is possible to make a supervised classification. The method of Denoeux is based on the use of k nearest neighbors. For this bioprocess, 68 samples were used: 30 samples for the state 1 (called class C_1), 7 samples for the state 2 (called class C_2), 31 samples for the state 3 (called class C_3). The approach of Denoeux [45] let us compute the masses for the classes C_1, C_2, C_3 and for the set Θ by using the following equations:

$$m(C_i) = \frac{m_i(C_i) \prod_{j \neq i} m_j(\Theta)}{K} \tag{15}$$

$$m(\Theta) = \frac{\prod_{i=1}^{3} m_i(\Theta)}{K} \tag{16}$$

where K is the following factor of normalization:

$$K = \sum_{i=1}^{3} \prod_{j \neq i} m_j(\Theta) + (1-3) \prod_{i=1}^{3} m_i(\Theta) \tag{17}$$

With:

$$m_i(C_i) = 1 - \prod_{x_{ki} \in C_i} (1 - \alpha_0 e^{-d^{ki,l}}) \tag{18}$$

$$m_i(\Theta) = \prod_{x_{ki} \in C_i} (1 - \alpha_0 e^{-d^{ki,l}}) \tag{19}$$

where $d^{ki,l}$ represents the metric distance between each element to classify x_l and the labeled sample x_{ki} of the class C_i ($i = \{1, 2, 3\}$) for each biochemical parameter; e represents the exponential function and α_0 is a fixed value between 0 and 1 (see [45]). For this application we have chosen $\alpha_0 = 0.95$. For the experimentation, we have tested several values of k for the k nearest neighbours (from $k = 1$ to $k = 7$).

As we use a method which computes the masses only for singletons (ie. here each of the 3 exclusive classes) and for the set Θ, the choices of the plausibility or the belief lead here to similar results in the classification.

6.3. Experimental results

First of all, let us clarify that in this subsection, when we talk about conflict of sources, it is in a general way (it is not the Dempster's conflict): it means conflict from the difference from distance mesure of belief function theory. And before we discuss about the results of classification, let us analyze the experimental interest of the relevance of the sources of information (which are here biochemical parameters). The relevance should have an experimental sense, otherwise it will not be useful. The following examples are given for $k = 7$ (number of k nearest neighbours), $\tau = 0.3$ and with

the measure based on norm 1 (but with the Jousselme distance the results are quite similar). For example in the figure 2, for the variable RQ (the Respiratory Quotient) the singularity localized at $5.35h$ (which is actually an artifact) is considered as not relevant by the method, and consequently is not taken into account during the classification. Thus the classification tends to make fault detection and eliminate those fault detections. Moreover, the parameter temperature which is a regulated parameter is relevant from t=0h to t=8.05h and then becomes irrelevant before the beginning of state 2 (see figure 3). When we analyse this parameter, we note that it doesn't change until t=15h where it is changed by the expert. The relevance comes again from t=16.5h until the death of the micro-organisms and corresponds to the change made by the expert (the time delay of the relevance may correspond to a lag time of the biological system).

We can see that the evaluation of the relevance provides consistent results concerning the regulated parameters. These results are confirmed by the experts in microbiology. This confirmation of the relevance by the expert in microbiology is not so surprising: actually the estimation of the relevance of the data sources is a non-model-based approach and it is based on the data from the process. As data of the bioprocess express the biological phenomenon (and also the biochemical actions of the expert) the estimation of relevance is thus biologically consistent.

Figure 2: A failure appears in the RQ at t=5.35h, but it is eliminated in the classification

Figure 3: The relevance of the parameter temperature disappears just before the beginning of the state 2 and appears again when the expert modifies the temperature (on this figure the values are normalised, x-axis is the time, y-axis is the value of parameters).

Secondly, with these first experimental results, we can make an empirical comparison between the Jousselme distance and the distance based on norm 1. We recall that they both belong to the Minkowski family of measures. Let us take some biochemical parameters to compare the value of their conflicts computed with the Jousselme distance and with the distance based on norm 1: we can see that the value of the average of each conflict is similar for the two distances. For example for the measured carbon dioxide (see figure 4), the two curves (representing the mean of the conflict with the Jousselme distance and with the distance based on norm 1) seems to be similar; the average of the conflict from the distance based on norm 1 seems to be a positive translation of the one from the Jousselme distance. In figure 5, results are similar for the acidity (pH). This is confirmed if we compute statistical properties of the average of the conflict of all biochemical parameters at each measured time. In figure 6, minimum, maximum, and average of the conflict from the distance based on norm 1 are respectively close to minimum, maximum and average of the conflict based on Jousselme distance. In fact the conflicts from the distance based on norm 1 are not positively translated from the ones from the Jousselme distance, but the differences between the derivatives are close to 0: more precisely, the difference between the two

derivatives is at the scale of 10^{-3} whereas the values of conflicts, whatever the biochemical parameter, are close to 10^{-1} (see table 1 and figure 7). Thus they have a similar trend. This is confirmed by the difference between the standard deviation which is also close to 0. Furthemore, the distance between the conflict from the conflict the Jousselme distance and the conflict from the distance based on norm 1 is at the scale of 10^{-2}(see for example the figure 8). We note that the conflict from the distance based on norm 1 is always higher, than the one from the Jousselme distance, whatever the biochemical parameter.

Thus the conflict have similar behavior but the conflict from the distance based on norm 1 is higher so more *restrictive* than the one from the Jousselme distance: at each time, conflict from distance from norm 1 will have less relevant biochemical parameters (relevant data sources) than the conflict from the Jousselme distance.

Figure 4: The two averages of conflict respectively from the distance based on norm 1 and from the Jousselme distance for the measured carbon dioxide. The first curve (conflict from distance based on norm 1) seems to be positively translated from the second (conflict from Jousselme distance).

Last but not least, with regards to the classification, the approach presented in this paper provides better results. We note that it is difficult to provide a benchmark and to compare this approach with other methods because, as we have seen above, each method of evaluation of the relevance of

Figure 5: The two averages of conflict respectively from the distance based on norm 1 and from the Jousselme distance for the measure of acidity (pH). The first curve (conflict from distance based on norm 1) seems to be positively translated from the second (conflict from Jousselme distance).

Figure 6: Maximum, average, minimum, and standard deviation of the average of the conflict respectively from distance based on norm 1 and from Jousselme distance.

Figure 7: Values at each time of the difference between the derivative of the average of the conflict based on norm 1 for all biochemical parameters and the derivative of the average of the conflict based on the Jousselme distance for all biochemical parameters.

Figure 8: Values at each time of the difference between the average of the conflict based on norm 1 for all biochemical parameters and the average of the conflict based on the Jousselme distance for all biochemical parameters.

differences	value
between derivatives of the 2 conflicts of pH (acidity)	5.1×10^{-3}
between derivatives of the 2 conflict of the average of all parameters	2.3×10^{-3}
between the standard deviations of the average of the 2 conflicts	6.8×10^{-3}

Table 1: Values of difference between conflicts from distance based on norm 1 and from Jousselme distance. First line is the average of the differences between the 2 averages of the 2 conflicts for the acidity. Second line is the average of the differences between the 2 averages of the 2 conflicts for all the biochemical parameters. Third line is the average of the differences between the 2 standard deviation of the average of the 2 conflicts for all the biochemical parameters.

data sources depends on the definition of relevance which has been chosen. However, it is possible to analyse the results of classification. As an expert in microbiology give us a normal behaviour, we can compare the results of the classification with and without estimation of relevance. The improvement come from 6 points to 20 points compared to the result of classification without relevance. We have tested the two measures, and the results are given in the table 2. We can see that, whatever the measure, whatever the values of threshold, the results are better with than without the estimation of relevance than without this estimation of relevance. Moreover, results with the two measures are quite similar, with a little advantage (but *a priori* not really significant) for the measure of conflict based on norm 1. This little difference can be explained by the fact that the measure of conflict based on norm 1 is more restrictive than the one based on Jousselme distance. Thus

the selected biochemical parameters, considered as relevant, are less numerous than those of the Jousselme distance and therefore give more precise information for the classification. However, we repeat that regardless of the chosen measure, this approach improves the perform of the classification.

We note that experimental results show that the threshold influences the results of correct classification. The optimization of the threshold τ regarding the results of classification will be part of further works. Moreover, as we said above, another approach [39] adjusting the relevance of data source (by using discounting) in order to satisfy a predefined level of conflict could also be used in this application. Another alternative method is the use of a weakened factor depending on the value of the average distance measure [37]; this could let us avoid the use of a direct threshold as τ . These approaches should be tested in future works.

In any case, the taking into account of the relevance improves the results of classification and this is, from our point of view, an important element of this work.

threshold τ	Jousselme distance	norm 1 distance
0,4	97%	99%
0,5	74%	88%
0,6	73%	75%
0,7	75%	75%
no estimation of relevance	69%	69%

Table 2: Percentage of good classification in function of the threshold τ.

7. Conclusion

We have proposed the use of belief functions theory for the evaluation of relevance of the sources of data and for the diagnosis by classification of bioprocesses. The diagnosis is viewed as the best choice of variables to determine the physiological states. The definition of the relevance that we use is based on the distance measures in the framework of belief functions theory. We compare two distance measures, and we show that the results of the classification are empirically similar. On the one hand, the estimations of relevance from this approach of the data source are confirmed by the expert in microbiology: this method provides meaningful and consistent information about the relevance of the data sources. On the other hand, whatever the chosen measure, the estimation of relevance improves the percentage of correct classification.These two results improve the process: relevance carries out using theory of believe function contributes to the optimization of bioprocesses. Further works should concern the analysis of the relevance on time sets (instead of an analysis of a discrete system) and tests using various optimizing approaches as mentionned above: the use of degree of reliability and the use of discounting.

Acknowledgment

We would like to thank the anonymous reviewers for their constructive comments which have greatly helped us to improve the paper. We also thank the staff of the INSA laboratory of biotechnology-bioprocesses of Toulouse

for their help and advices and particularly to Prof. Jean-Louis Urribelarea . This work is partially supported by Directorate of Youth, Sports and Social Cohesion of Guadeloupe, French Caribbean.

References

[1] J. Steyer, Sur une approche qualitative des systèmes physiques : aide en temps réèl à la conduite des procédés fermentaires, Thse de Doctorat, Université Paul Sabatier, Toulouse France (Décembre 1991).

[2] A. Doncescu, J. Waissman, G. Richard, G. Roux, Characterization of bio-chemical signals by inductive logic programming, Knowledge-Based Systems 15 (1-2) (2002) 129–137.

[3] S. Régis, J. Desachy, A. Doncescu, Evaluation of biochemical sources pertinence in classification of cell's physiological states by evidence theory, in: FUZZ'IEEE, Budapest, Hongrie, 2004.

[4] S. Régis, A. Doncescu, J. Desachy, Théorie des fonctions de croyance pour la fusion et l'évaluation de la pertinence des sources d'informations: application à un bioprocédé fermentaire, Traitement du signal 24 (2) (2007) 115–132.

[5] A.-L. Jousselme, D. Grenier, E. Bossé, A new distance between two bodies of evidence, Information Fusion 2 (2001) 91–101.

[6] S. Régis, L. Faure, A. Doncescu, J.-L. Uribelarrea, L. Manyri, J. Aguilar-Martin, Adaptive physiological states classification in fed-batch fermentation process, in: IFAC CAB'9, Nancy, France, 2004.

[7] S. Régis, A. Doncescu, Knowledge based discovery in fed-batch bioprocess, in: IFAC CAB'10, Cancun, Mexico, 2007.

[8] S. Régis, A. Doncescu, J. Desachy, Detection and characterization of physiological states in bioprocesses based on hölder exponent, Knowledge-Based Systems 21 (2008) 70–79.

[9] R. Greiner, A. Grove, A. Kogan, Knowing what doesn't matter: exploiting the omission of irrelevant data, Artificial Intelligence 97 (1997) 345–380.

[10] M. Lazo-Cortès, J. Ruiz-Schulcloper, Determining the feature relevance for non-classically described objects and a new algorithm to compute typical fuzzy testors, Pattern Recognition Letters 16 (1995) 1259–1265.

[11] A. Blum, P. Langley, Selection of relevant features and examples in machine learning, Artificial Intelligence 97 (1997) 245–271.

[12] S. Baluja, D. Pomerleau, Dynamic relevance: vision-based focus attention using artificial neural networks, Artificial Intelligence 97 (1997) 381–395.

[13] L. Zadeh, A note on web intelligence, world knowledge and fuzzy logic, Data and Knowledge Engineering 50 (2004) 291–304.

[14] G. Paltoglou, M. Salampasis, M. Satratzemi, Modeling information sources as integrals for effective and efficient source selection, Information Processing and Management 47 (2011) 18–36.

[15] A. Spoerri, Examining the authority and ranking effects as the result list depth used in data fusion is varied, Information Processing and Management 43 (2007) 1044–1058.

[16] F. Pichon, D. Dubois, T. Denoeux, Relevance and truthfulness in information correction and fusion, International Journal of Approximate Reasoning, 2011 (doi:10.1016/j.ijar.2011.02.006).

[17] The Merriam-Webster dictionary, http://www.merriam-webster.com/dictionary, visit date: February 1st 2008.

[18] le Trésor de la Langue Française Informatisé, http://atilf.atilf.fr/tlf.htm, CNRS-ATILF.

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

[20] X.-W. Chen, An improved branch and bound algorithm for feature selection, Pattern Recognition Letters 24 (2003) 1925–1933.

[21] R. Felix, Relationships between goals in multiple attribute decision making, Fuzzy Sets and Systems 67 (1994) 47–52.

[22] A. Dempster, A generalisation of bayesian inference, Journal of the Royal Statistical Society 30 (1968) 205–247.

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

[24] R. Yager, On the Dempster-Shafer framework and new combination rules, Information Sciences 41 (1987) 93–138.

[25] P. Smets, Non standard Logics for Automated Reasoning, Academic Press, 1988, Ch. Belief Functions, pp. 29–39.

[26] P. Smets, R. Kennes, The transferable belief model, Artificial Intelligence 66 (1994) 191–234.

[27] E. Lefevre, O. Colot, P. Vannoorenberghe, Belief function combination and conflict management, Information Fusion 3 (2002) 149–162.

[28] J. Schubert, Analyzing the combination of conflicting belief functions, Information Fusion (8) (2007) 387–412.

[29] L. Zadeh, A mathematical theory of evidence (book review), AI magazine 5 (3) (1984) 81–83.

[30] R. Haenni, Shedding new light on zadeh's criticism of dempster's rule of combination, in: Proceedings of the Eighth International Conference on Information Fusion, USA, 2005, pp. 879–884.

[31] J. Schubert, On non specific evidence, International Journal of Intelligent Systems 8 (1993) 711–725.

[32] J. Schubert, Specifying nonspecific evidence, International Journal of Intelligent systems 11 (1996) 525–563.

[33] W. Liu, Analyzing the degree of conflict among belief functions, Artificial Intelligence 11 (170).

[34] J. Schubert, The internal conflict of a belief function, in: T. Denoeux, M.-H. Masson (Eds.), Belief Functions: Theory and Applications, Proceedings of the Second International Conference on Belief Functions, Springer, Berlin, 2012, pp. 169–177.

[35] A. Martin, About conflict in the theory of belief functions, in: Belief Functions, France, 2012, pp. 161–168.

[36] S. Destercke, T. Burger, Revisiting the notion of conflicting belief functions, in: Belief Functions, France, 2012, pp. 153–160.

[37] A. Martin, A.-L. Jousselme, C. Osswald, Conflict measure for the discounting operation on belief functions, in: Information Fusion, Germany, 2008.

[38] A.-L. Jousselme, P. Maupin, Distances in evidence theory: Comprehensive survey and generalizations, International Journal of Approximate Reasoning 53 (2) (2012) 118–145.

[39] J. Schubert, Conflict management in dempster-shafer theory using the degree of falsity, International Journal of Approximate Reasoning (52) (2011) 449–460.

[40] J. Schubert, Finding a posterior domain probability distribution by specifying nonspecific evidence, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 3 (1995) 163–185.

[41] X. Li, J. Dezert, X. Huang, Selection of sources as a prerequesite for information fusion with application to slam, in: Fusion 06, Florence, Italy, 2006.

[42] M. Chebbah, A. Martin, B. Yaghlane, Modélisation dans les bases de données évidentielles, EGC-AFDC 1O, Hammamet.

[43] F. Cuzzolin, A geometric approach to the theory of evidence, IEEE Transactions on Systems, Man, and Cybernetics - Part C (38) (2008) 522–534.

[44] S. Régis, A. Doncescu, J. Desachy, Estimation of relevance and fusion of data sources using belief function theory: application to bioprocess, in: CSTST 08, Paris, France, 2008.

[45] T. Denoeux, A k-nearest neighbor classification rule based on Dempster-Shafer theory, IEEE trans. on systems, man, and cybernetics 25 (5) (1995) 804–813.