

## *Chapitre 4*

# *Simulation et Evaluation des Performances*

## INTRODUCTION

Ce dernier chapitre est consacré à la simulation et l'application des techniques de reconnaissance de formes proposées comme étant une solution pour la surveillance de la qualité de l'eau. L'objectif est de valider et d'évaluer les performances de ces méthodes présentées. Les exigences principales d'efficacité sont formulées sur deux points essentiels à savoir, les tests de spécification qui vérifient que le programme réalise bien la tâche pour laquelle il a été conçu, et les tests de performances qui vont servir à mesurer l'efficacité avec laquelle cette tâche est remplie. On évaluera pour les méthodes exposées les paramètres liés au taux de reconnaissance, au temps d'apprentissage et à l'erreur d'entraînement. Une discussion des résultats conclura cette étude de simulation.

### 1. Problématique

#### 1.1. Architecture du système de contrôle et de surveillance

Il s'agit dans cette partie de travail d'évaluer les performances de la technique choisie qui est issue, rappelons-le, du domaine de l'intelligence artificielle à savoir, les SVM. Des techniques servant comme outils de base pour l'aide à la décision et présentant une réponse plus élaborée par rapport aux autres techniques se basant sur des données brutes, venant directement des variables de surveillance, ou à partir de données traitées venant des sorties de traitements de bas niveau. Le choix effectué sur la base des résultats obtenus, conduira à l'intégration de la technique sélectionnée au niveau d'un système de surveillance assurant un contrôle permanent de la qualité de l'eau. L'architecture de ce système imaginé est basée sur une approche multi-capteurs et présentée dans la figure 4.1. Le processus de contrôle est vu comme un problème de reconnaissance de formes, où les classes correspondent aux différents états de l'eau, et les formes représentent l'ensemble des observations ou mesures des paramètres liés à ses caractéristiques.

Au niveau du système, on peut supposer que les différents paramètres physico-chimiques utilisées, tels que le pH, la température ( $T^\circ$ ), la conductivité (C), la turbidité (TU), etc sont transformés en signaux électriques à partir des capteurs physiques, et transmis vers une station de contrôle qui assure l'acquisition, le traitement et l'analyse. La technique de surveillance utilisée effectuée après chaque acquisition, la classification et la séparation des données en plusieurs classes bien différentes. Une suite d'acquisitions pourrait être envisagée plusieurs fois par jour, sous des conditions prédéfinies. Un module d'apprentissage supervisé par un expert, permet de collecter de manière continue les paramètres relatifs aux différents états de l'eau pour la mise en œuvre d'une base de connaissance complète.

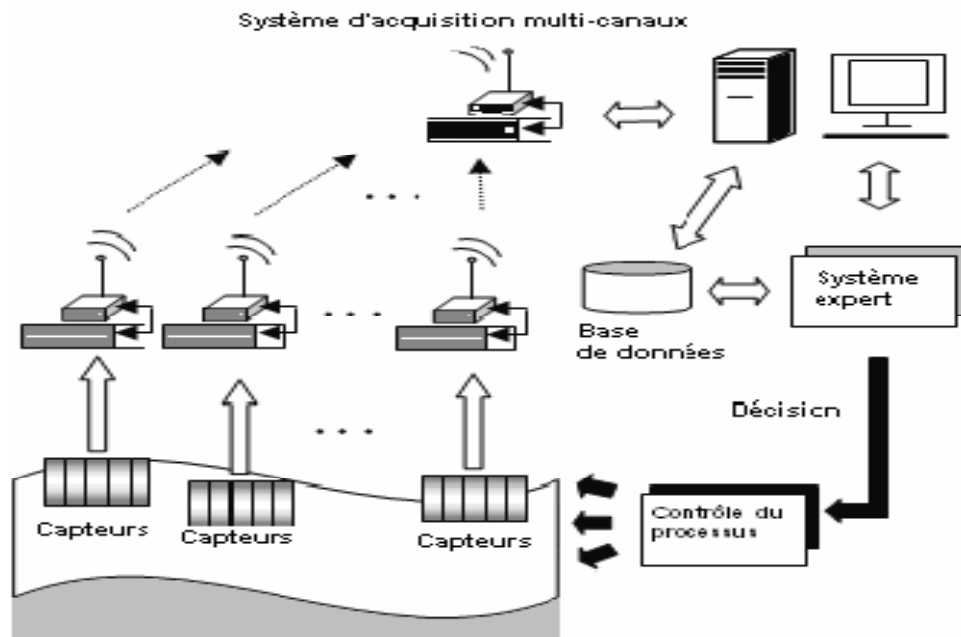


Figure 4.1. Architecture du système de contrôle et de surveillance.

## 1.2. Approche utilisée dans la surveillance

La solution devant être adoptée par la technique citée ci-dessus au problème de reconnaissance de formes posé, ne s'applique en fait que si on se trouve dans le cas d'un apprentissage supervisé. Nous procédons donc lors d'une étape préliminaire d'apprentissage, à paramétrer le classificateur pour la reconnaissance. L'étape de test ou de reconnaissance proprement dite, s'effectue une fois le modèle statistique établi. Il y a ici tout l'intérêt pour dire que cette approche se caractérise par sa souplesse et sa généricité. A souligner toutefois que les méthodes de reconnaissance de formes à base d'apprentissage statistique sont les plus utilisées dans les systèmes de classification à fusion multi-sensorielle. En général l'apprentissage est une étape assez longue, et nécessite plus de temps de calcul. Les techniques partagent ce point commun mais diffèrent sur un certain nombre d'autres points. L'étude effectuée dans les paragraphes suivants en fera la différence. Ce critère (temps d'apprentissage) aussi important dans le choix du modèle de reconnaissance, évoque un traitement hors ligne devant être effectué par le système de surveillance. Le déroulement de cette opération en permanence contribue sans doute à enrichir une base de connaissance qu'on veut qu'elle soit la plus complète possible pour le modèle de surveillance implanté. Le système de contrôle doit donc pouvoir marier à la fois une surveillance directe de l'eau et un apprentissage en arrière plan (en différé). Un opérateur (ou système) expert supervisant cet

apprentissage permet de collecter de manière continue les paramètres relatifs aux différents états de l'eau.

## 2. Description des données d'entrée

Nous cherchons à décider sur la qualité de l'eau à travers ses paramètres descripteurs. Nous n'avons en fait aucune connaissance a priori sur un type de modèle représentant parfaitement ce procédé, par contre nous pouvons porter notre jugement sur la qualité de cette eau à partir de quelques données descriptives. Il y a quatre paramètres physico-chimiques qui sont souvent utilisés dans plusieurs travaux [50], et qui renseignent sur les dangers majeurs qu'il faut surveiller. Ces paramètres sont résumés comme suit : Conductivité (C), pH, Température ( $T^\circ$ ), et Turbidité (TU).

L'objectif qui se trouve derrière la collecte des données relatives à ces paramètres est de trouver un modèle de classification permettant de distinguer trois états bien distincts de l'eau (Classe I, Classe II et Classe III). La qualité de cette eau reflétée par sa potabilité repose en fait sur une corrélation qui ne peut être identifiée que statistiquement. Des données descriptives expérimentales recueillies sur une période de trois ans (2009-2011) à partir de la station de Tilesdit (Bouira – Algérie) comme étant une zone d'étude dans ce travail qui pourraient atteindre cet objectif. A noter que la turbidité et le pH sont fortement dépendants des phénomènes saisonniers. Il y a donc intérêt de disposer d'au moins une année pour archiver des données afin de déterminer une base de connaissance assez complète capable de fonctionner normalement. D'où la nécessité d'une base de connaissance riche en informations exigeant d'abord une collecte des données sur une longue période, et la présence d'un expert.

## 3. Application à la station de production d'eau potable TILES DIT

### 3.1. Le site TILES DIT

Le barrage « Tilesdit » est situé géographiquement dans la commune de Bechloul à 20 km du Sud-Est de la wilaya de Bouira, Algérie. Ce barrage est situé entre les coordonnées cartographiques et les coordonnées Lambert suivantes (figure 4.2) :

- Latitude :  $35^\circ 13' 22''$  Nord.
- Longitude :  $4^\circ 14' 23''$  Est.



Figure. 4.2. Carte géographique situant le barrage « Tilesdit » [Google Maps].

Ce barrage disposant d'un volume de 167 million de mètres cubes d'eau, est conçu de façon à juguler la tension qui persiste dans la distribution d'eau au niveau de 12 communes (figure 4.3). De part son implantation dans la région de Bouira, le barrage Tilesdit dont la mise en eau a été effectuée vers la fin de l'année 2005, garantira de l'eau potable aux communes relevant de l'Est et du Sud-Est de la wilaya, c.-à-d: El-Asnam, Bechloul, El-Adjiba, Ahl-Ksour, Oud El-Berdi, Mesdour, Taguedit, Ahnif, Bordj O'khris, Ouled rached, Ath Mansour, Bouira et la zone industrielle de Sidi Khaled. Le transfert d'eau dont le lancement des travaux de réalisation a été prévu au début 2011, est destinée à l'alimentation de cinq autres communes de la daïra de Mansourah, dans la wilaya de Bordj Bou Arréridj. Des travaux sont en cours en vue de raccorder les communes de Takdit et Ait Laaziz, ainsi que d'autres communes rurales du Sud-Est de la wilaya de Bouira au réseau d'alimentation de ce barrage, qui devrait s'étendre jusqu'à Hammam k'sana. Parallèlement, la même direction a prévu 12 opérations portant sur la réalisation de réseaux AEP, en plus de 12 réservoirs d'une capacité globale de  $74000 \text{ m}^3$  d'eau en vue de l'amélioration de l'approvisionnement en eau des communes alimentées par ce barrage notamment avec un débit de  $72650 \text{ m}^3/\text{jour}$ . Il est également important de signaler qu'à l'horizon 2020, ce barrage garantira l'eau en faveur de 307200 habitants, selon les prévisions du secteur de l'hydraulique. Les travaux menés ont porté à 6 le nombre de stations de pompage pour un débit de 41 à 981 l/s. Une station de traitement d'une capacité de  $74000 \text{ m}^3/\text{jour}$ , un réservoir d'eau traitée de deux compartiments d'un volume total de  $13000 \text{ m}^3$  et de 6 réservoirs de capacité allant de 200 à  $5000 \text{ m}^3$ .



Figure .4.3. Image représentant le site du barrage « Tilesdit » [Google Earth].

### 3.2. La station de production d'eau potable TILESdit

L'eau prélevée dans le barrage est pompée jusqu'à la station de traitement. Celle-ci se trouvant au même lieu, est mise en service depuis 2009 figure 4.4. Elle effectue le processus d'épuration à travers les cinq étapes de traitement indiquées au premier chapitre à savoir : le prétraitement, la pré-oxydation, la clarification, la désinfection et l'affinage. L'étape de clarification est assurée par le procédé de coagulation-floculation, décantation et filtration, grâce à un décanteur et un étage de filtration sur sable.



Figure. 4.4. Image représentant le site de la station « Tilesdit » [Google Earth].

### 3.3. Prétraitement des données

#### 3.3.1. Données d'entrée

Nous cherchons à appliquer notre approche de surveillance aux paramètres descripteurs de la qualité de l'eau brute fournis par les capteurs de mesure de la station. Notre connaissance du processus de traitement est limitée aux données enregistrées de la station durant les trois années écoulées 2009-2011[43]. Ces mesures proviennent des différentes étapes de traitement, on y trouve :

- des mesures en continu issues de capteurs physico-chimiques,

- des analyses faites au laboratoire.

Plusieurs paramètres descripteurs de la qualité de l'eau brute mesurés en ligne quotidiennement à raison de 3 fois/jour, en plus des essais effectués au laboratoire qui sont réalisés chaque semaine. Quatre paramètres descripteurs principaux sont mesurés directement des capteurs vers la station sont : Température, pH, Conductivité et Turbidité. Ces paramètres sont mesurés en continu et à tout niveau du processus de traitement (Eau brute, Eau décantée, Eau filtrée et Eau traitée). D'autres sont aussi mesurés quotidiennement tels que l'Ammonium et le Nitrite. Les paramètres tels que : Calcium, Magnésium, Chlorure, Sulfate, Bicarbonate, Dureté Total TH, Dureté Permanente, Titre Alcalin et Titre Alcalin Complet, sont mesurés une fois par semaine. La Couleur est par contre mesurée une fois/jour à tous les niveaux de traitement. Le Chlore Résiduel libre est ainsi mesuré chaque jour au niveau des étapes de décantation-filtration et à la sortie de la station (Eau traitée).

En conclusion, quatre paramètres descripteurs sélectionnés par cette analyse préliminaire (Température, Conductivité, pH et Turbidité) peuvent donc être retenus pour la surveillance. Ce sont les mêmes qui sont mesurés de façon permanente à tous les niveaux de traitement de la station.

## **4. Technique de contrôle et de surveillance**

### **4.1. SVM multi-classe**

La méthode SVM a été largement utilisée pour résoudre des problèmes de classification dans plusieurs domaines d'application. La classification est la technique de contrôle utilisée pour la surveillance de la qualité de l'eau dans cette application.

Dans ce qui suit, nous présentons la combinaison de la méthode ACP avec la technique SVM, et ce toujours dans le but du contrôle et de surveillance de la qualité de l'eau. L'ACP est utilisée pour réduire la dimension des variables d'entrée pour en extraire celles qui sont pertinentes et non redondantes. La technique SVM reste plutôt utilisée et appliquée pour le même objectif, à savoir, la classification. Pour cette situation, il s'agit d'une classification multi-classe où la sortie peut appartenir à trois classes bien différentes (classe I: Très bonne, classe II: moyenne, classe III: médiocre). L'intérêt d'une telle application, est de montrer la validité de la conception basée sur une réduction de variables en entrée même pour une classification plus complexe. Si l'idée vient d'être confirmée, les résultats obtenus pourraient évidemment être généralisés pour le cas bi-classe. Pour entamer la résolution de ce problème, l'une des solutions proposées consiste à diviser celui-ci en un ensemble de classifications

binaires avant de les combiner. La méthode SVM étendue peut être imaginée comme solution. L'approche *un-contre-tous* représente une stratégie très populaires pour les SVM multi-classes. Les performances de classification demeurent tout de même liées au choix de la fonction noyau et de ses paramètres.

## 4.2. Approches de classification

### 4.2.1. Principe

Comme on le sait les SVM sont développées pour traiter essentiellement des problèmes binaires, mais elles peuvent être adaptées pour traiter des problèmes de type multi-classe. Il y a eu pour cela plusieurs tentatives combinant des classifieurs binaires pour cerner ce problème. Nous allons dans ce qui suit expliquer brièvement la méthode *Un contre Tous* parmi les méthodes les plus utilisées :

- **Approche : Un contre Tous**

L'approche "*un contre tous*" (OAA) est la plus simple et la plus ancienne des méthodes de décomposition permettant d'aborder un problème de discrimination à catégories multiples (cas multi-classe). Elle consiste à utiliser autant de classifieurs binaires (à valeurs réelles) par catégorie que de classes. Chaque classifieur renvoie la valeur 1 si la forme à reconnaître appartient à la classe, -1 sinon. Le  $k^{ième}$  classifieur est destiné à distinguer la catégorie d'indice  $k$  de toutes les autres. Pour affecter un exemple, on le présente donc à  $Q$  classifieurs, et la décision s'obtient en application du principe "winner-takes-all" : l'étiquette retenue est celle associée au classifieur ayant renvoyé la valeur la plus élevée. Il faut donc pour reconnaître une forme, la soumettre à tous les classifieurs, le meilleur est celui qui remporte la décision. Il est évident qu'avec un nombre de classes élevé, la combinatoire peut devenir énorme. Cette méthode suppose donc la construction de  $N$  classifieurs et  $N$  comparaisons pour la décision.

En résumé, l'idée consiste simplement à transformer le problème à  $k$  classes en  $k$  classifieurs binaires. Le classement est donné par le classifieur qui répond le mieux.

### 4.2.2. Optimisation de SVM par PSO

Faut-il souligner que les paramètres  $\sigma$  et  $C$  associés au noyau RBF influent souvent sur la classification [41]. Dans cette section, nous décrivons notre système PSO-SVM, conçu à optimiser la précision du classifieur SVM c'est-à-dire la détermination des paramètres libres : « le coefficient de régularisation  $C$  et le sigma  $\sigma$  » en se basant sur une méthode évolutionnaire telle que l'optimisation par essaim de particules. L'application du PSO de l'optimum global dans le problème de détermination des paramètres libres  $\sigma$  et  $C$  consiste à minimiser les écarts entre les valeurs de sortie désirée  $y_d$  et les valeurs de sortie calculée  $y_c$ .

Ceci consiste à trouver le minimum de l'erreur quadratique moyenne de généralisation (EQMG) suivant :

$$EQMG = \frac{1}{N} \sum_{i=1}^N (y_{di} - y_{ci})^2 \quad (4.1)$$

où  $N$  est le nombre d'exemples de la base d'apprentissage.

La méthode PSO commence par une initialisation aléatoire de particules dans l'espace de recherche. A chaque itération de l'algorithme, PSO choisie la meilleure particule à partir de la population entière. Ce processus est répété jusqu'à ce que le SVM converge vers ses meilleures performances.

## 5. Simulation et Evaluation

Comme nous l'avons vu, l'entraînement d'une SVM consiste à résoudre un problème d'optimisation quadratique convexe. Le choix de la méthode à utiliser est critique car les performances de l'implantation en seront directement tributaires. Notre sélection s'est portée sur une méthode à points intérieurs appelée : IPM (Interior Points Method). Cette méthode semble donner de très bons résultats en termes de temps de calcul et de précision de la solution [45]. L'implantation d'IPM est basée sur le package d'optimisation LOQO [44]. Celui-ci permet de traiter des problèmes quadratiques plus généraux, il est considéré comme le plus efficace.

### 5.1. Apprentissage et Test

Pour mener notre application de surveillance, un ensemble d'apprentissage et de test est constitué à partir d'une base de données de 800 échantillons correspondant aux 4 paramètres descripteurs retenus. Pour cette situation, il s'agit d'une classification multi-classe où la sortie peut appartenir à trois classes bien différentes (classe I: Très bonne, classe II: moyenne, classe III: médiocre). A signaler toutefois que cette base de données est réalisée selon les normes de potabilité recommandées [46, 47, 48]. L'ensemble de ces données est séparé en deux : 400 échantillons pour l'apprentissage et 400 pour le test.

La technique SVM est appliquée pour effectuer une classification multiclasse en utilisant l'approche OAA. Dans le tableau 4.1, on montre les résultats obtenus avec l'usage d'une fonction noyau RBF et les paramètres  $\sigma$  et  $C$  associés au noyau RBF calculés par PSO.

Tableau. 4.1. Classification utilisant le modèle SVM.

Variables d'entrée	Noyau Base de données	Approches multi-classes	Taux de classification (%)		Temps d'apprentissage (s)
			Apprentissage	Test	
(4 paramètres) (C, pH, T°, TU) 400/400	RBF ( $\sigma = 2.19$ , C=1)	OAA	100 %	69.75%	51.7
	RBF ( $\sigma = 2.19$ , C=10)		100 %	68.25%	51.7
	RBF ( $\sigma = 2.19$ , C=0.1)		<b>100 %</b>	<b>95.5%</b>	<b>104.5</b>

Tableau. 4.2. Classification utilisant le modèle PSO-SVM.

Variables d'entrée	Noyau Base de données	Approches multi-classes	Taux de classification (%)		Temps d'apprentissage (s)
			Apprentissage	Test	
(4 paramètres) (C, pH, T°, TU) 400/400	RBF ( $\sigma = 0.5$ , C=1000)	OAA	<b>100 %</b>	<b>95.5 %</b>	<b>51.7</b>

On remarque que le taux de classification ainsi obtenu est supérieur de 68% dans tous les cas. Un taux de 95.5 % est obtenu dans l'utilisation du modèle PSO-SVM avec le noyau RBF. Ce phénomène peut être expliqué par l'efficacité de la PSO dans le calcul des paramètres  $\sigma$  et  $C$ .

## 5.2. Discussion des résultats

Pour la technique SVM, la maximisation de la marge séparatrice entre les trois classes est bien maîtrisée. Nous avons trouvé que cette méthode fournit de très bons résultats pour le cas d'une classification multi-class. Les solutions trouvées dépendent exclusivement des exemples (base de données) présentés en entrée. La technique naturellement basée sur l'apprentissage par exemples. On peut dire d'après les résultats obtenus, que la fonction noyau RBF soit standard et unique pour toutes les bases de données utilisées. Une erreur d'entraînement nulle (satisfaction du principe MRS), ainsi qu'un taux de reconnaissance plus de 95.5 % confirment clairement l'adéquation de la technique avec ce type d'application. L'enrichissement continu de cette base ainsi que la présence d'un expert, est donc valorisant pour cette technique.

Un contrôle de potabilité étendu à une grande échelle, peut être pris en charge de façon dynamique par le système de la figure 4.1, puisque le temps de calcul réalisé est très faible. Ce modèle est donc retenu pour être appliqué en classification.

On résume dans le tableau 4.3 un état des caractéristiques liées aux solutions envisageables dans l'utilisation de modèle SVM dans un système de classification.

Tableau. 4.3. Tableau des caractéristiques de modèle SVM.

Propriétés	SVM
Algorithme	Apprentissage et généralisation
Principe	MRS
Base de données	Toute la base de données
Optimisation	Quadratique (hessienne) non linéaire
Apprentissage	Par exemples (base de données)
Architecture de réseau	Standard
Ajustement des poids	Stable
Les poids	Suivant une formulation mathématique (problème dual, multiplicateurs de Lagrange)
Classification des données	Binaire et multi-classe
Paramètres d'apprentissage	Moins de paramètres
Séparation des données	La possibilité de maximiser la marge séparatrice. (Contrôle de classification)
Méthode	L'utilisation des noyaux (probabilité augmente)
Temps d'apprentissage	Long
Taux de reconnaissance	>95 %
Inconvénient majeur	Choix des paramètres de noyau – Solution par PSO

### 5.3. Principales caractéristiques

Les résultats caractéristiques correspondant aux deux modèles étudiés sont résumés dans le tableau 4.4 :

Tableau. 4.4. Caractéristiques principales des modèles (SVM et PSO-SVM).

Caractéristiques	SVM	PSO-SVM
Base de données		
Temps d'apprentissage (sec)	104.5 s	51.7 s
Erreur d'apprentissage	0	0
Taux de reconnaissance (%)	95.5%	95.5%

Il apparait clairement que les deux modèles étudiés présentent en général de très bons résultats, avec des taux de reconnaissance acceptables sur le plan décisionnel. L'effet de

l'utilisation de la technique PSO pour la sélection optimale des paramètres  $\sigma$  et  $C$  sur les performances de classification a une légère influence apparemment, des taux de classification similaires sont obtenus. Par conséquent une diminution en matière de temps d'apprentissage est bien évidente, ce qui se voit clairement en comparant les tableaux 4.1 et 4.2. Cette optimisation par PSO a engendré un effet positif, à savoir, un maintien des performances de classification et parfois meilleur, une recherche directe des paramètres d'apprentissage et le temps d'apprentissage diminue.

Le modèle PSO-SVM est plutôt mieux placé du point de vue temps d'apprentissage, ce qui lui confère l'avantage d'une intégration dans un système de contrôle dynamique. L'enrichissement de la base de données peut faire une amélioration en matière de taux de reconnaissance, ce qui laisse envisager son intégration dans un système de surveillance en continu permettant la collecte en permanence des données supervisées par un expert. Un contrôle de potabilité étendu à une grande échelle, peut être pris en charge de façon dynamique par le système de la figure 4.1, puisque le temps de calcul réalisé est très faible.

Etant donné que la technique PSO-SVM est retenue comme étant le meilleur choix, elle est appliquée dans ce cadre pour effectuer une classification multi-classe en utilisant les approches OAA. L'impact de ce résultat est important sur les plans, aussi bien technique (recherche directe des paramètres et temps d'apprentissage plus faible), qu'économique (nombre réduit de capteurs). Les caractéristiques affichées dans les tableaux 4.3 et 4.4 soulignent l'intérêt théorique et pratique du modèle PSO-SVM pour ce type d'application. Ce modèle est donc retenu pour être appliqué en classification.

## CONCLUSION

Ce quatrième et dernier chapitre a fait l'objet d'une étude en simulation concernant la mise en œuvre de la technique SVM dans le domaine du contrôle et de surveillance des eaux propres. Cette étude a permis la validation et l'évaluation des performances cette méthode présentée. Les paramètres liés au taux de reconnaissance, au temps d'apprentissage et l'erreur d'entraînement, ont été les facteurs pertinents qui ont permis d'évaluer la méthode étudiée. L'utilisation de la méthode PSO se révèle efficace pour les problèmes d'optimisation non linéaires et semble très performant en termes de la précision des solutions trouvées. Pour notre système, nous avons opté pour l'algorithme PSO de l'optimum global car il semble bien adapté au problème posé en termes de temps de calcul. La discussion des résultats obtenus, a permis d'opter notre choix pour l'approche proposée pour ses qualités et avantages adaptés au

problème posé. Présentée pour un problème de classification multi-classe, la technique hybride a fourni de très bons résultats de simulation.