ⵜⴰⴳⵍⴷⵉⵜ ⵜⴰⵎⴰⵣⵉⵖⵜ ⵜⴰⵣⴰⵢⵔⵉⵜ ⵜⴰⵏⴰⵙⵙⵉⵔⵜ ⵜⴰⵏⴰⵎⵓⵔⵜ
République Algérienne Démocratique et Populaire
ⵜⴰⵙⴷⵓⵙⵜ ⴰⵏⴰⵎⵓⵔ ⵜⴰⵙⵙⵉⵔⵜ ⴰⵏⴰⵎⵓⵔ ⵜⴰⵎⵓⵔⵜ
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

N° Réf :……

Centre Universitaire

Abd elhafid Boussouf Mila

Institut des sciences et de la technologie **Département de Mathématiques et Informatiques**

Mémoire préparé en vue de l'obtention du diplôme de Master

En : Informatique

Spécialité : Sciences et Technologies de l'information et de la communication (STIC)

Les modèles hybrides basés optimisation et apprentissage automatique pour la résolution de problèmes dans le domaine de la bioinformatique

Préparé par :

Maamar Radja

Merrouche Saliha

Soutenue devant le jury

Encadré par : Mme Afri Faiza M.A.B

Président : Mme Zekiouk Mounira M.A.B

Examinateur : Mlle Bouchekouf Asma M.A.A

Année universitaire : 2015/2016

Remerciement

Nous Remercions en tout premier lieu ALLAH le tout puissant, qui nous'a donné la force, la volonté et lecourage pour accomplir ce modeste travail.

Nous avons abouti un travail, qui a été le résultat d'un cheminement de tout un parcours pédagogique, qui a duré tout le long de notre parcours éducatif dansl'enseignement supérieur.

Un remerciement particulier à notre encadreur ***Mme Afri Faiza*** *pour sa présence, son aide et surtout pour ses précieux conseils qui nous ont assistés pour l'accomplissement de notre projet.*

Nos vifs remerciements sont également aux membres du jury pour l'intérêt qui 'ils ont porté à notre travail en acceptant de l'examiner et de l'enrichir par leurs propositions et commentaires.

Nous tenons à exprimer nos sincères remerciements à tout le personnel de l'institut des sciences et de la technologie surtout les enseignants qui nous ont formé durant toutes nos années d'étude.

Un remerciement particulier à nos très chers parents, frères, sœurs, collègues et amies respectives qui nous ont encouragés, soutenu durant tout notre parcours.

MERCI À TOUS *Maamar Radja*

Merrouche Saliha

Dédicace

Chaque fois qu'on achève une étape importante dans notre vie, on fait une pose pour regarder en arrière et se rappeler toutes ces personnes qui ont partagé avec nous tous les bons moments de notre existence, mais surtout les mauvais.

Je dédie ce modeste travail en signe de reconnaissance et de respect.

-A mon père «Cherif».

-A ma très chère mère Naima qui m'a fait protéger pendant toute ma vie.

-A mes frères Samir et Abdalkawi.

-A mes sœurs Mounira, Samia, Salma et Ikhlass a pour leur soutien et encouragement.

-A ma chère enseignante Mme Afri Faiza et mon binôme saliha.

-A mes chère amies : MERIEM, SARA, Saliha, Rokiya et Hadjer,…

-A tous mes collègues surtout Lokmane, Hicham et Amir.

-A tous la famille Maamar et la Famille Iyadi.

-A tous mes enseignants sans exception.

RADJA

Dédicace

Chaque fois qu'on achève une étape importante dans notre vie, on fait une pose pour regarder en arrière et se rappeler toutes ces personnes qui ont partagé avec nous tous les bons moments de notre existence, mais surtout les mauvais.

Je dédie cet humble travail :

-A mon père «Abdelhani», et à ma mère «Houria» qui m'ont donné la tendresse et l'amour.

-A tous mes frères Bilal, Hamza, Mohamed, sa3dan, Younes, Chouaib, Islam.

-A mon Fiancé Yousef pour son soutien et encouragement.

-A ma belle sœur Samira et ma tante Malika

-A tous mes oncles et tantes, Surtout oncle « Farouk », ainsi que mes cousins, surtout « Hatem ».

-A mon binômes Radja, qui a été mon bras droit, ma source d'inspiration et qui m'a soutenu tout au long de ce travail.

-A mes chère amies : MERIEM, FATIMA, AMINA, AMEL

-A toute la famille Merrouche et la Famille Kihal;

-A tous ceux qui ont aidé de près ou de loin à Réaliser ce travail.

SALIHA

Résumé. L'identification des biomarqueurs devient l'un des sujets de recherche scientifique les plus abordés. Ce n'est pas une tâche facile en raison de l'énorme volume de données «omiques». Néanmoins, il peut être modélisé comme le problème de la sélection d'attributs. Les techniques de sélection d'attributs ont montrés leur efficacité dans l'identification de biomarqueurs pour le cancer. Dans des travaux récents, l'hybridation de l'apprentissage automatique et de l'optimisation a démontré sa puissance dans diverses applications. Dans ce travail, nous proposons une approche hybride pour identifier des biomarqueurs à partir des données d'expression génique. Nous avons combiné PSO et la recherche Cuckoo améliorée ICS avec le classifieur SVM pour sélectionner des biomarqueurs. Dans la méthode proposée, les meilleures particules sélectionnées par PSO sont l'entrée de l'algorithme de ICS. Des résultats expérimentaux sur des ensembles de données de puces à ADN, ont prouvé que cette hybridation mène à des résultats satisfaisants.

Mots clés: intelligence par essaim, la recherché coucou (CS), optimization par essaim de particules (PSO), machines a vecteurs de support (SVM), decouverte de biomarqueurs, selection d'attributs, hybridation.

Abstract. Identifying biomarkers becomes one of the surge scientific research matters. It is not an easy task due to the huge volume of "Omics" data. Nonetheless, it can be modeled as the problem of selecting features. Feature selection techniques shown their efficiency in identifying biomarkers for cancer. In recent works, the hybridization of machine learning and optimization demonstrated its power in various applications. In this paper, we propose a hybrid approach to identify biomarkers from gene expression data. We combined the swarm intelligence PSO and Cuckoo search with SVM classifier to select biomarkers with best classification performance. In the proposed method, the best particles selected by PSO are the input of Cuckoo search algorithm. Experimental results on DNA microarray datasets proved that this hybridization gives sufficient results.

Key words: Swarm intelligence, Cuckoo search (CS), Particle swarm optimization (PSO), Support vector machines (SVM), Biomarker discovery, Feature selection, hybridization.

ملخص .تحديد المؤشرات الحيوية أصبحت واحدة من أكثر المواضيع التي نوقشت في البحث العلمي .هذه ليست مهمة سهلة بسبب الكم الهائل من البيانات البيولوجية .ومع ذلك، فإنه يمكن أن على غرار باعتبارها مشكلة اختيار الصفات .وقد أظهرت تقنيات اختيار السمات كفاءتها في تحديد المؤشرات الحيوية للسرطان .في الأعمال الأخيرة، وقد ثبت ان التهجين وتحسين قوتها في مختلف التطبيقات .في هذا العمل، نقترح نهجا الهجين لتحديد المؤشرات الحيوية من البيانات التعبير الجيني . دمجنا PSOو ICS مع SVM لتحديد المؤشرات الحيوية .في الطريقة المقترحة، فإن أفضل الجسيمات الناتجة عن PSO هي المدخلات الخوارزمية ل .ICSوقد أثبتت نتائج التجارب على مجموعات البيانات DNA microarray أن هذا التهجين يؤدي إلى نتائج مرضية.

كلمات مفتاحية: سرب الاستخبارات، سعى الوقواق(CS) ، سرب الجسيمات(PSO) ، شعاع الدعم الآلي (SVM) اكتشاف المؤشرات الحيوية، اختيار الصفات،التهجين.

Table des matières

Liste des Figures :

Liste des Tables :

Liste des Algorithmes :

Introduction Générale

Introduction générale :

Contexte de l’étude :

La bioinformatique est un domaine scientifique. Elle consiste en l'utilisation de méthodes et d'outils informatiques pour le traitement massif de l'information biologique. Avec l'apparition de technologies de biologie moléculaire de plus en plus complexes, on voit aujourd'hui émerger des projets où collaborent biologistes et informaticiens, et souvent aussi chimistes, physiciens, mathématiciens…etc. Le bioinformaticien doit être capable d'interagir avec ces différents spécialistes.

En raison des difficultés d’interprétation biologique d’un grand nombre de gènes. Nous avons pu montrer qu'il est possible de détecter un petit nombre de gènes qui sont impliqués de manière quasi-certaine dans l'irradiation, et qu’ils permettent de construire un système de classification très performant.

La recherche d’un sous ensemble d’attributs optimal à partir d’un ensemble d’attributs de dimension élevée est un problème d’optimisation NP-difficile. Pour facilite la sélection de ces données, réduire le temps de calcul, et pour des résultats plus efficace, les informaticiens proposent l’hybridation des algorithmes d’apprentissage avec les algorithmes d’optimisation.

Problématique et motivation :

La bioinformatique se situe à l'interface de deux disciplines : la **biologie** et l'**informatique**. Le bio-informaticien peut donc aisément être comparé à un interprète bilingue connaissant à la fois le langage informatique et le langage biologique. Il doit pouvoir se servir de l'outil informatique afin de résoudre des problèmes d'ordre biologique faisant appel à des données biologiques !

Actuellement, la taille de ces bases de données croît exponentiellement, notamment en raison du séquençage des génômes, il est donc nécessaire d'avoir une technique efficace pour exploiter ces données.

Le processus connu sous le nom de « Sélection d'attributs » a pour but de filtrer le vecteur des attributs de base, de sorte que l'espace de caractéristiques soit réduit de façon optimale selon certains critères d'évaluation.

Dans ce mémoire, nous nous concentrons sur le problème de sélection d'attributs dans le domaine de la bioinformatique, c'est pour ça nous avons proposé un algorithme hybride PSO-ICS-SVM basé sur l'algorithme d'optimisation par essaim particulaire (PSO) ; l'algorithme de recherche coucou amélioré (ICS), et l'algorithme Machines à Vecteurs Supports (SVM).

Structure du mémoire :

Le travail présenté dans ce mémoire s'intègre dans le domaine de la sélection des attributs, Ce mémoire est structuré en quatre chapitres :

-Dans le premier chapitre, nous allons présenter quelques notions de base sur l'apprentissage automatique et l'optimisation, ainsi quelques algorithmes.

-Dans le deuxième chapitre, nous allons présenter ce que la bioinformatique, ensuite nous exposerons la découverte et les étapes du développement des biomarqueurs, et à la fin nous terminerons par le processus de la sélection d'attributs.

-Dans le troisième chapitre, nous allons exposer la notion d'hybridation et ses stratégies et nous présentons quelques travaux connexes, et à la fin nous allons mener une étude générale sur l'algorithme d'optimisation par essaim particulaire (PSO) ; l'algorithme de recherche coucou amélioré(ICS), et nous exposerons l'algorithme machines à vecteurs de support (SVM).

- Dans le quatrième chapitre, nous présenterons le Matériel, le langage de programmation utilisé, ainsi les données qui ont été utilisées pour valider l'approche proposée. Les tests utilisés, et les résultats obtenus sont présentés.

Chapitre I :

L'Apprentissage Automatique et l'Optimisation

I. Introduction :

L'apprentissage automatique consiste à utiliser des ordinateurs pour optimiser un modèle de traitement de l'information selon certains critères de performance à partir d'observations, que ce soit des données-exemples ou des expériences passées.

D'autre part, l'optimisation est une branche des mathématiques qui permet de résoudre des problèmes en déterminant le meilleur élément d'un ensemble selon certains critères prédéfinis. De ce fait, l'optimisation est omniprésente dans tous les domaines et évolue sans cesse depuis Euclide.

Dans ce chapitre, nous allons présenter les notions de base sur l'apprentissage automatique et l'optimisation, nous précisons les algorithmes de chacune.

II. Apprentissage automatique :

1. Définition :

L'apprentissage automatique (machine learning), est un champ d'étude de l'intelligence artificielle, concerne la conception, l'analyse, le développement et l'implémentation de méthodes permettant à une machine d'évoluer par un processus systématique, et ainsi de remplir des tâches difficiles ou impossibles à remplir par des moyens algorithmiques plus classiques.[1]

2. Applications :

L'apprentissage automatique est utilisé pour doter des ordinateurs ou des machines de systèmes de perception de leur environnement : vision, reconnaissance ce d'objets (visages, schémas, langages naturels, écriture, formes syntaxique, etc.) ; moteurs de recherche ; aide aux diagnostics, médical notamment, bioinformatique, chémoinformatique ; interfaces cerveau machine ; détection de fraudes à la carte de crédit, analyse financière, dont l'analyse du marché boursier ; jeu ; génie logiciel ; sites Web adaptatifs ou mieux adaptés ; locomotion de robots ; etc.[1]

3. Types d'apprentissage :

Les algorithmes d'apprentissage peuvent se catégoriser selon le mode d'apprentissage qu'ils emploient :

Apprentissage supervisé :

Si les classes sont prédéterminées et les exemples connus, le système apprend à classer selon un modèle de classement ; on parle alors d'apprentissage supervisé (ou d'analyse discriminante).Un expert (ou oracle) doit préalablement correctement étiqueter des exemples. L'apprenant peut alors trouver ou approximer la fonction qui permet d'affecter la bonne « étiquette » à ces exemples. Parfois il est préférable d'associer une donnée non pas à une classe unique, mais une probabilité d'appartenance à chacune des classes prédéterminées (on parle alors d'apprentissage supervisé probabiliste). L'analyse discriminante linéaire ou les SVM sont des exemples typiques. Autre exemple : en fonction de points communs détectés avec les symptômes d'autres patients connus (les « exemples »), le système peut catégoriser de nouveaux patients au vu de leurs analyses médicales en risque estimé (probabilité) de développer telle ou telle maladie.il faut alors modéliser ce système.[1]

Modélisation mathématique

y = h(x /o)

h(o) : fonction générale du modèle

o:parametres du modèle

Objectif : apprendre une projection entre des observations X en entrée et des valeurs associées Y en sortie

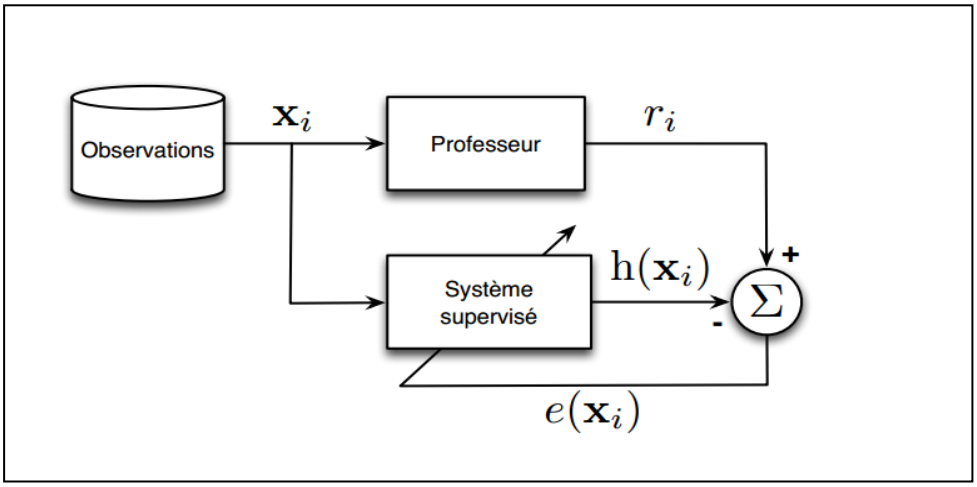

Figure 1 : Schématisation de l'apprentissage supervise

Classement :

Y ; est discret et correspond à des étiquettes de classes ;

H() :est une fonction discriminante.

L'Apprentissage non-supervisé :

Quand le système ou l'opérateur ne dispose que d'exemples, mais non d'étiqueté, et que le nombre de classes et leur nature n'ont pas été prédéterminés, on parle d'apprentissage non supervisé (ou clustering). Aucun expert n'est disponible ni requis. L'algorithme doit découvrir par lui-même la structure plus ou moins cachée des données. Le système doit ici dans l'espace de description (la somme des données) cibler les données selon leurs attributs disponibles, pour les classer en groupe homogènes d'exemples. La similarité est généralement calculée selon la fonction de distance entre paires d'exemples. C'est ensuite à l'opérateur d'associer ou déduire du sens pour chaque groupe. Divers outils mathématiques et logiciels peuvent l'aider. [1]

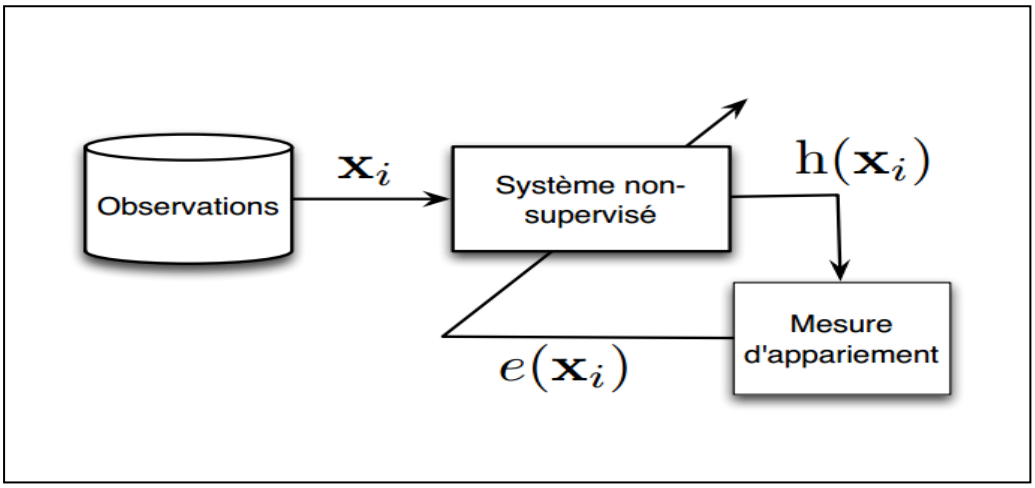

Figure 2 : Schématisation de l'apprentissage supervisé

L'apprentissage semi-supervisé :

Effectué de manière probabiliste ou non, il vise à faire apparaitre la distribution sous-jacente des « exemples » dans leur espace de description. Il est mis en œuvre quand des données (ou « étiquettes») manquent... Le modèle doit utiliser des exemples non-étiquetés pouvant néanmoins renseigner. [1]

L'apprentissage partiellement supervisé (probabiliste ou non) :

Quand l'étiquetage des données est partiel. C'est le cas quand un modèle énonce qu'une donnée n'appartient pas à une classe A, mais peut-être à une classe B ou C (A, B et C étant 3 maladies par exemple évoquées dans le cadre d'un diagnostic différentiel).[1]

L'apprentissage par renforcement :

L'algorithme apprend un comportement étant donné une observation. L'action de l'algorithme sur l'environnement produit une valeur de retour qui guide l'algorithme d'apprentissage.[1]

L'apprentissage par transfert :

L'apprentissage par transfert peut être vu comme la capacité d'un système à reconnaître et appliquer des connaissances et des compétences, apprises à partir de tâches antérieures, sur de nouvelles tâches ou domaines partageant des similitudes. La question qui se pose est : comment identifier les similitudes entre la ou les tâche(s) cible(s) et la ou les tâche(s) source(s), puis comment transférer la connaissance de la ou des tâche(s) source(s) vers la ou les tâche(s) cible(s). [1]

4. Les algorithmes utilisés :

Les algorithmes d'apprentissages de type supervisé

- **Les arbres de décision :**

Un arbre de décision est un outil d'aide à la décision représentant un ensemble de choix sous la forme graphique d'un arbre. Les différentes décisions possibles sont situées aux extrémités des branches (les « feuilles » de l'arbre), et sont atteints en fonction de décisions prises à chaque étape. L'arbre de décision est un outil utilisé dans des domaines variés tels que la sécurité, la fouille de données, la médecine, etc.

Un avantage majeur des arbres de décision est qu'ils peuvent être calculés automatiquement à partir de bases de données par des algorithmes d'apprentissage supervisé. Ces algorithmes sélectionnent automatiquement les variables discriminantes à partir de données non-structurées et potentiellement volumineuses. Ils peuvent ainsi permettre d'extraire des règles logiques de cause à effet (des déterminismes) qui n'apparaissaient pas initialement dans les données brutes.[2]

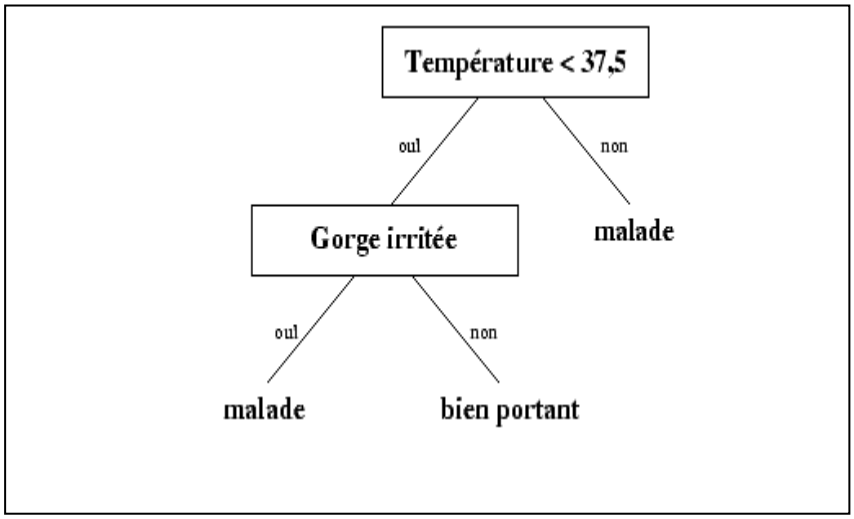

Figure 3 : Hiérarchie de l'arbre de décision [2]

- **La méthode des k plus proches :**

En intelligence artificielle, la méthode des k plus proches voisins est une méthode d'apprentissage supervisé. En abrégé k-NN ou KNN, de l'anglais k-nearest neighbor.

Dans ce cadre, on dispose d'une base de données d'apprentissage constituée de N couples « entrée-sortie ». Pour estimer la sortie associée à une nouvelle entrée x, la méthode des k plus proches voisins consiste à prendre en compte (de façon identique) les k échantillons d'apprentissage dont l'entrée est la plus proche de la nouvelle entrée x, selon une distance à définir.

Par exemple, dans un problème de classification, on retiendra la classe la plus représentée parmi les k sorties associées aux k entrées les plus proches de la nouvelle entrée x.[3]

- **Les réseaux de neurones**

Un réseau de neurones artificiels est un modèle de calcul dont la conception est très schématiquement inspirée du fonctionnement des neurones biologiques.

Les réseaux de neurones sont généralement optimisés par des méthodes d'apprentissage de type probabiliste, en particulier bayésien. Ils sont placés d'une part dans la famille des applications statistiques, qu'ils enrichissent avec un ensemble de paradigmes permettant de créer des classifications rapides (réseaux de Kohonen en particulier), et d'autre part dans la famille des méthodes de l'intelligence artificielle auxquelles ils fournissent un mécanisme

perceptif indépendant des idées propres de l'implémenteur, et fournissant des informations d'entrée au raisonnement logique formel.

En modélisation des circuits biologiques, ils permettent de tester quelques hypothèses fonctionnelles issues de la neurophysiologie, ou encore les conséquences de ces hypothèses pour les comparer au réel.[4]

- **Un modèle de mélange gaussien :**

Est un modèle statistique exprimé selon une densité mélange. Il sert usuellement à estimer paramétriquement la distribution de variables aléatoires en les modélisant comme une somme de plusieurs gaussiennes (appelées noyaux). Il s'agit alors de déterminer la variance, la moyenne et l'amplitude de chaque gaussienne. Ces paramètres sont optimisés selon un critère de maximum de vraisemblance pour approcher le plus possible la distribution recherchée. Cette procédure se fait le plus souvent itérativement via l'algorithme espérance-maximisation (EM).

Les modèles de mélange gaussien sont réputés reconstruire de manière particulièrement efficace les données manquantes dans un jeu de données expérimentales. [5]

- **L'analyse discriminante linéaire :**

Fait partie des techniques d'analyse discriminante prédictive. Il s'agit d'expliquer et de prédire l'appartenance d'un individu à une classe (groupe) prédéfinie à partir de ses caractéristiques mesurées à l'aide de variables prédictives.

L'analyse discriminante linéaire peut être comparée aux méthodes supervisées développées en apprentissage automatique et à la régression logistique développée en statistique. [6]

- **Les machines à vecteurs de support :**

(En anglais Support Vector Machine, SVM) sont un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de discrimination et de régression. Les SVM sont une généralisation des classifieur linéaires.

Les SVM ont été développés dans les années 1990 à partir des considérations théoriques de Vladimir Vapnik sur le développement d'une théorie statistique de l'apprentissage : la théorie de Vapnik-Chervonenkis. Les SVM ont rapidement été adoptés pour leur capacité à travailler avec des données de grandes dimensions, le faible nombre d'hyper paramètres, leurs garanties théoriques, et leurs bons résultats en pratique.

Les SVM ont été appliqués à de très nombreux domaines (bio-informatique, recherche d'information, vision par ordinateur, finance…). Selon les données, la performance des machines à vecteurs de support est de même ordre, ou même supérieure, à celle d'un réseau de neurones ou d'un modèle de mélanges gaussiens.[7]

Les algorithmes d'apprentissage de type non supervisé

- **La logique floue** :

Est une extension de la logique booléenne créée par Lotfi Zadeh en 1965 en se basant sur sa théorie mathématique des ensembles flous, qui est une généralisation de la théorie des ensembles classiques. En introduisant la notion de degré dans la vérification d'une condition, permettant ainsi à une condition d'être dans un autre état que vrai ou faux, la logique floue confère une flexibilité très appréciable aux raisonnements qui l'utilisent, ce qui rend possible la prise en compte des imprécisions et d'incertitudes.

- **Le partitionnement k-moyennes (clustering):**

Le partitionnement en k-moyennes (ou k-means en anglais) est une méthode de partitionnement de données et un problème d'optimisation combinatoire. Étant donnés des points et un entier k, le problème est de diviser les points en k partitions, souvent appelés clusters, de façon à minimiser une certaine fonction. On considère la distance d'un point à la moyenne des points de son cluster ; la fonction à minimiser est la somme des carrés de ces distances.

Il existe une heuristique classique pour ce problème, souvent appelée méthodes des k-moyennes, utilisée pour la plupart des applications. Le problème est aussi étudié comme un problème d'optimisation classique, avec par exemple des algorithmes d'approximation.

Les k-moyennes sont notamment utilisées en apprentissage non supervisé où l'on divise des observations en k partitions. [30]

- **Cartes auto adaptatives :**

Cartes auto-organisatrices ou cartes topologiques forment une classe de réseau de neurones artificiels fondée sur des méthodes d'apprentissage non-supervisées.

Elles sont souvent désignées par le terme anglais self organizing maps (SOM), ou encore cartes de Kohonen du nom du statisticien ayant développé le concept en 1984. La littérature utilise aussi les dénominations : Réseau de Kohonen, Réseau auto-adaptatif ou Réseau auto-organisé.

Elles sont utilisées pour cartographier un espace réel, c'est-à-dire pour étudier la répartition de données dans un espace à grande dimension. En pratique, cette cartographie peut servir à réaliser des tâches de discrétisation, quantification vectorielle ou classification. [31]

5. Les facteurs de pertinence et d'efficacité d'apprentissage :

La qualité de l'apprentissage et de l'analyse dépendent du besoin en amont et a priori de la compétence de l'opérateur pour préparer l'analyse. Elle dépend aussi de la complexité du modèle (spécifique ou généraliste), de son adéquation et de son adaptation au sujet à traiter. Enfin, la qualité du travail dépendra aussi du mode (de mise en évidence visuelle) des résultats pour l'utilisateur final (un résultat pertinent pourrait être caché dans un schéma trop complexe, ou mal mis en évidence par une représentation graphique inappropriée).

- Nombre d'exemples (moins il y en a, plus l'analyse est difficile, mais plus il y en a, plus le besoin de mémoire informatique est élevé et plus longue est l'analyse) ;
- Nombre et qualité des attributs décrivant ces exemples. La distance entre deux « exemples » numériques (prix, taille, poids, intensité lumineuse, intensité de bruit, etc) est facile à établir, celle entre deux attributs catégoriels (couleur, beauté, utilité…) est plus délicate ;
- Pourcentage de données renseignées et manquantes ;
- « Bruit » : le nombre et la « localisation » des valeurs douteuses (erreurs potentielles, valeurs aberrantes…) ou naturellement non-conformes au pattern de distribution générale des « exemples » sur leur espace de distribution impacteront sur la qualité de l'analyse.[8]

III. l'Optimisation :

1. Définition :

L'optimisation c'est l'art de comprendre un problème réel, de pouvoir le transformer en un modèle mathématique que l'on peut étudier afin d'en extraire les propriétés structurelles et de caractériser les solutions du problème. Enfin, c'est l'art d'exploiter cette caractérisation afin de déterminer des algorithmes qui les calculent mais aussi de mettre en évidence les limites sur l'efficience et l'efficacité de ces algorithmes.[Paschos 05].

La résolution d'un problème d'optimisation nécessite l'étude de trois points particuliers :

- l'analyse du problème, regroupant la caractérisation du problème puis sa modélisation, qui aboutit en général à sa représentation informatique ou codage (définition et codage de l'ensemble des solutions réalisables).
- L'expression de l'objectif à optimiser, qui nécessite une bonne connaissance du problème.
- Le choix de la méthode de résolution, qui permet à partir de la représentation du problème dans un espace de recherche d'obtenir une solution ou un ensemble de solutions optimales par rapport à la fonction d'évaluation.

Ces schéma suivant est une illustration d'un framework général utilisé pour le passage d'un phénomène naturel observé à un algorithme inspiré de la nature, qui sera utilisé pour résoudre les problèmes d'optimisation.[9]

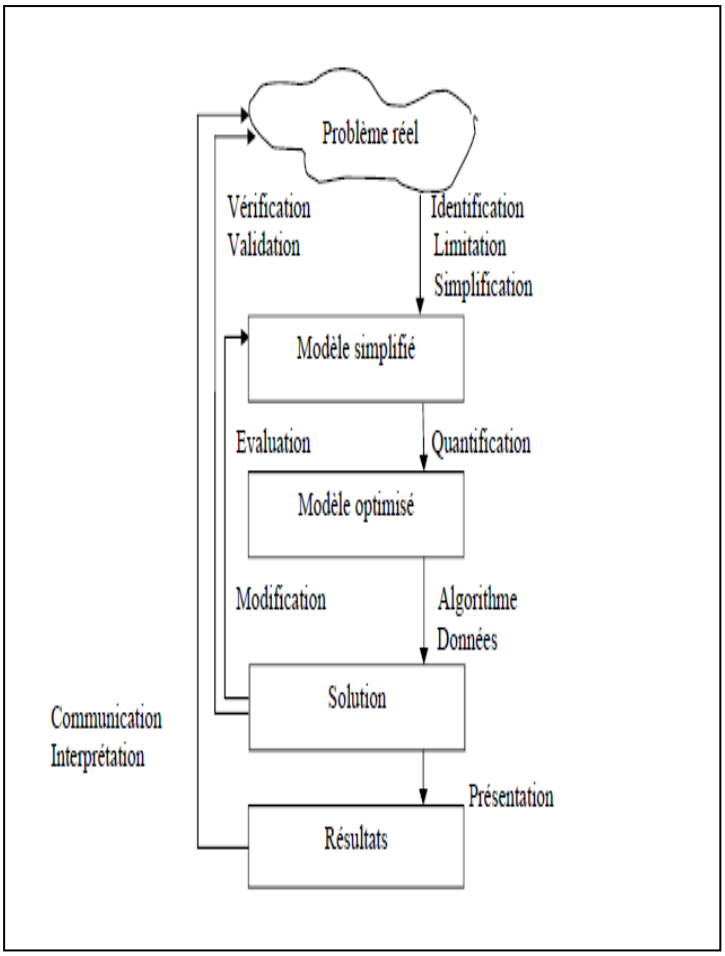

Figure 4 : Processus d'optimisation [Hazem et Janice 12]

2. Formulation mathématique :

$$\begin{cases} \text{Minimiser} : f_i(x), \quad (i = 1,2,\dots,M) \\ \quad x \in R^n (1.1) \\ \text{soumise à} : \emptyset_j(x) = 0, \quad (j = 1,2,\dots,J), \\ \quad \Psi_k(x) \leq 0, \quad (k = 1,2,\dots,K), \\ \quad \text{avec } x = (x_1, x_2, \dots, x_n) \end{cases}$$

x est un vecteur à éléments réels ou entiers de dimension n. Les composantes xi du vecteur x sont appelées variables de décision, elles peuvent être continues ou discrètes.

Les valeurs des xi représentent l'espace de rechercheR^n. La fonction fi(x) avec (i=1,2,…,M) est appelée la fonction de coût, la fonction économique ou la fonction objectif. Elle attribue à chaque instanciation de l'espace de recherche une valeur. M représente le nombre de fonctions objectif du problème Les inégalités $\emptyset_j$ et Ψ_k sont appelées les contraintes duproblème, qui peuvent être linéaires ou non linéaires. Pour un problème de minimisation, l'objectif est de rechercher les vecteurs qui minimisent la fonction f tout en respectant les contraintes

représentées par les fonctionsΨet$\emptyset$. Il est à noter que pour les problèmes demaximisation, il suffit de multiplier la fonction coût par (-1).[9]

3. Types d'optimisation :

Dans [Xin 10], plusieurs critères sont employés pour classifier les différents problèmes d'optimisation :

1- Classification basée sur le nombre d'objectifs : dans cette classification, on définit deux catégories de problèmes d'optimisation : ceux à un seul objectif (M=1) et ceux à plusieurs objectifs (M>1), on parle dans ce cas de problèmes d'optimisation multiobjectif. Il est important de souligner que la majorité des problèmes réels sont multiobjectifs.

2- Classification selon les contraintes : deux catégories sont constatées :

- problèmes sans contraintes (J=K=0).

- problèmes avec contraintes d'égalité (J≥1 et K=0), contraintes d'inégalité (J=0 etK≥1), ou avec les deux types de contraintes (J≥1, K≥1)

3- Classification en terme de la forme de fonction : si les fonctions de contraintes $\emptyset$j (x) et Ψk (x), ainsi que la fonction objectif fi(x) sont linéaires, on parle d'un problème d'optimisation linéaire. Dans le cas contraire on parle d'un problème d'optimisation non linéaire.

4- Classification selon l'allure de la fonction objectif qui peut admettre un seul optimum local qui est aussi l'optimum global, dans ce cas on parle d'un problème d'optimisation unimodal, par contre si la fonction objectif admet plusieurs optimums on parle d'un problème multimodal, par exemple la fonction f(x,y) = x2+y2 admet (0,0) comme minimum global, par contre la fonction

f(x,y) = sin(x)sin(y) est multimodale.

5- Classification selon le type des variables de décision : on parle de problèmes d'optimisation discrets ou combinatoires lorsque les variables de décision sont discrètes.

Parmi ces problèmes on peut citer le problème de voyageur de commerce, le problème de la coloration des graphes et le problème des N-reines. Par contre si les variables de décision sont réelles ou continues on parle dans ce cas des problèmes d'optimisation continue.

6- D'autre part, si les valeurs des variables de décision et de la fonction objective sont définies de manière exacte, on parle de problème d'optimisation déterministe. Mais en réalité quelques paramètres seulement sont connus et avec incertitude, le problème devient alors stochastique.

Généralement, on dit qu'un algorithme est stochastique si son comportement est déterminé par l'entrée mais aussi par des valeurs produites par un générateur de nombres aléatoires. Plusieurs exécutions successives de tels algorithmes ne produisent pas forcément le même résultat. La plupart des algorithmes stochastiques consistent à explorer l'espace des solutions, passant de l'une à l'autre suivant une logique propre à chacun .la figure 1.5 illustré la classification des problèmes. [9]

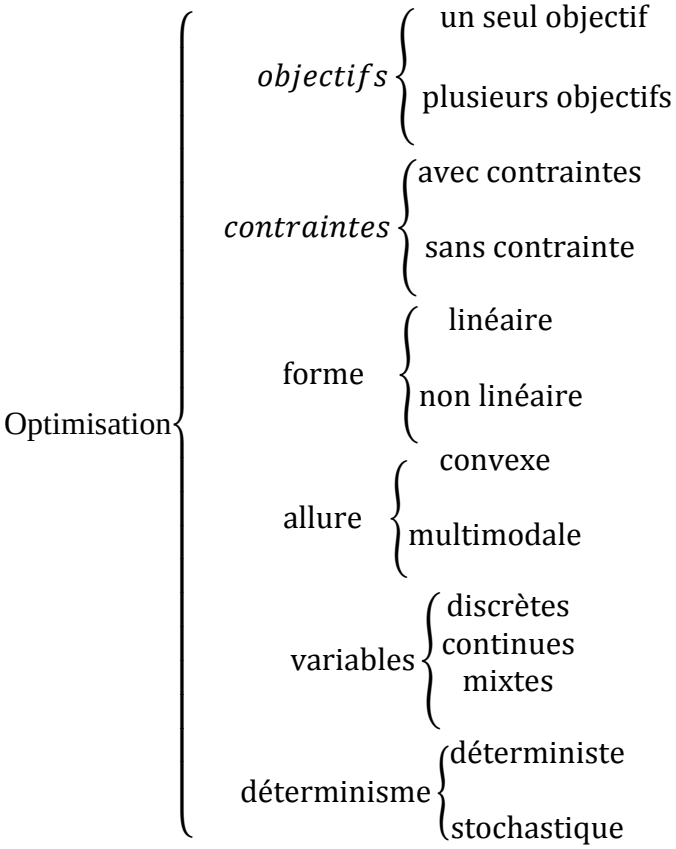

Figure 5 : Classification des problèmes d'optimisation [Xin 10]

4. Optimisation combinatoire :

L'optimisation combinatoire regroupe une large classe de problèmes d'optimisation ayant des applications dans de nombreux domaines aussi variés que la gestion, l'ingénierie, la production, les télécommunications, les transports, l'énergie et les sciences sociales.[9]

- **Définition :**

Un problème d'optimisation combinatoire consiste à parcourir l'espace de recherche afin d'en extraire une solution optimal parmi un ensemble fini de solutions d'une taille souvent très grand tel que son énumération exhaustive est une tache fastidieuse .

La résolution d'un problème d'optimisation combinatoire nécessite l'utilisation d'un procédé algorithmique permettant la maximisation ou la minimisation d'une ou de plusieurs fonctions (objectif) on respectant les contraintes posées par le problème. [9]

- **La complexité d'un problème :**

On entend ici par « complexité d'un problème » une estimation du nombre d'instructions à exécuter pour résoudre les instances de ce problème, cette estimation étant un ordre de grandeur par rapport à la taille de l'instance.

- **Classes de complexité :**

La notion de classe de complexité permet de classer les problèmes en fonction leur complexité. Les trois classes de complexité de problème les plus courantes sont P, NP, NP complet [Morelle 2010]:

Classe P : problèmes de décision pour lesquels on connaît des Algorithmes polynomiaux.

Classe NP : problèmes de décision pour lesquels on connaît des Algorithmes non déterministes polynomiaux (problèmes de décision pour lesquels n'importe quel certificat peut être vérifié en temps polynomial pour une réponse "oui").

NP-complet : un problème de décision A dans NP est NP-complet si tous les autres problèmes de la classe NP se transforment polynomialement dans le problème A.[10]

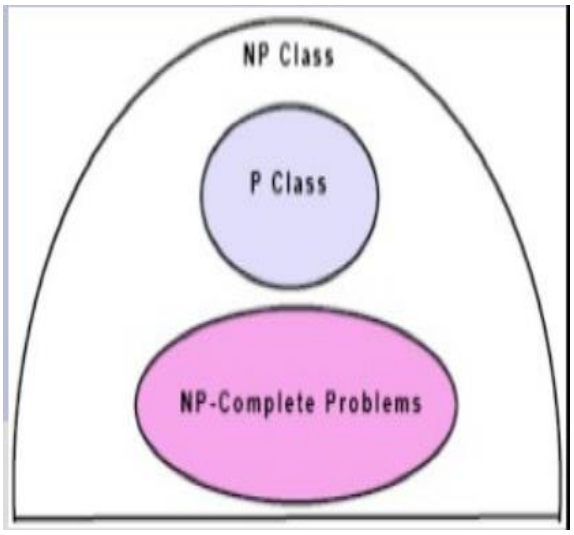

Figure 6 : Relation entre les ensembles P, NP, NP-Complet[10]

5. Méthodes d'optimisation :

Il existe 5 méthodes d'optimisation, citées ci-dessous :

-Méthode complète : trouve toujours une solution.

-Méthode optimale : trouve toujours la meilleure solution (optimum global).

-Méthode exacte (exhaustive) : explore l'espace de recherche dans sa totalité (énumération intelligente) ➔ optimale.

-Méthode approché (approximative) : explore une sous-partie de l'espace de recherche.

-Méthode déterministe : exécuté toujours la même suite d'opération.

-Méthode probabiliste (ou stochastique) : fait des choix probabilistes guidés par des tirages aléatoires.[11]

6. Classification des méthodes d'optimisation combinatoire

Les méthodes exactes ne sont efficaces que pour les instances de problèmes de petite taille,

On ne s'intéressera ici qu'aux méthodes approchées.

La figure 7 illustré quelques méthodes d'optimisation

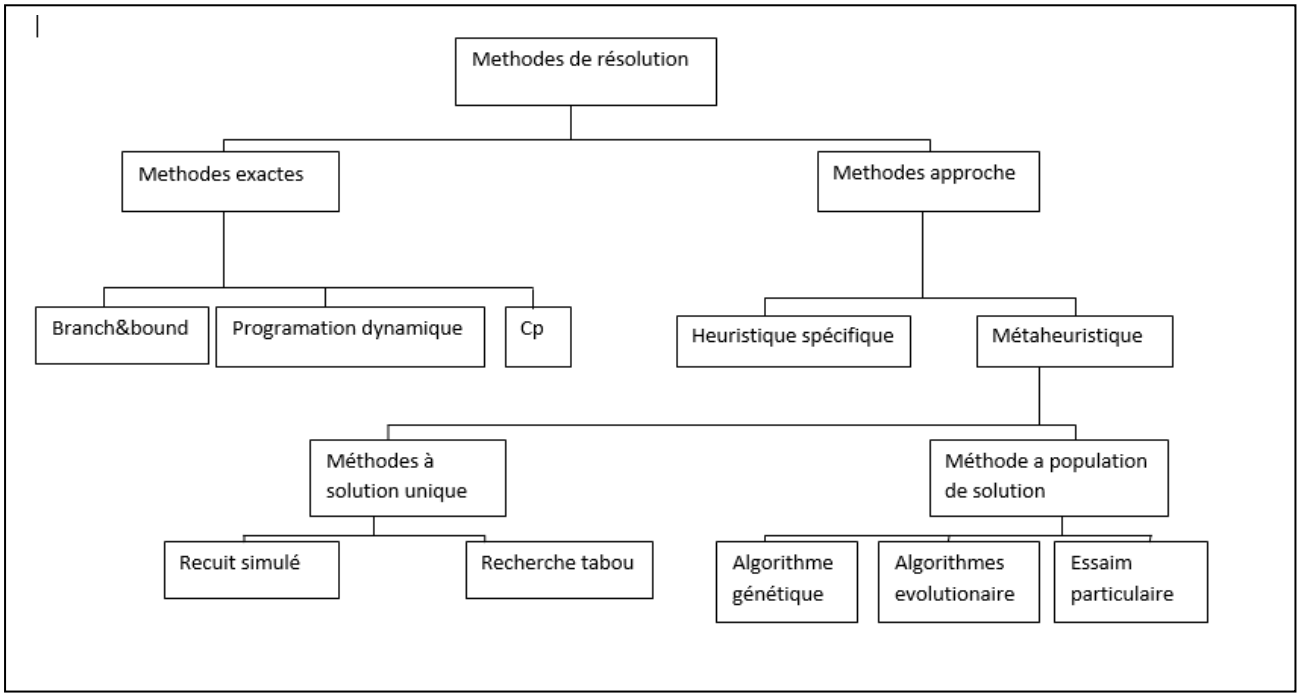

Figure 7 : Classification des méthodes d'optimisation combinatoire [11]

Les méthodes exactes :

Les méthodes exactes cherchent à trouver de manière certaine la solution optimale en examinant de manière explicite ou implicite la totalité de l'espace de recherche. Elles ont l'avantage de garantir la solution optimale néanmoins le temps de calcul nécessaire pour atteindre cette solution peut devenir très excessif en fonction de la taille du problème (explosion combinatoire) et le nombre d'objectifs à optimiser. Ce qui limite l'utilisation de ce type de méthode aux problèmes bi-objectifs de petites tailles. On peut citer quelque unes de Ces méthodes génériques : programmations dynamique, branch and bound, cut and Price (CP).[12]

Branch&Bound :

C'est une méthode générique de résolution exacte de problèmes d'optimisation, et plus particulièrement d'optimisation combinatoire. C'est une méthode constructive de recherche arborescente qui utilise l'énumération implicite basée sur la notion de bornes afin d'éviter l'énumération de larges classes de mauvaises solutions. Elle utilise la stratégie diviser pour régner, en se basant sur deux concepts : le

branchement (séparation) qui consiste à partitionner ou diviser l'espace des solutions en sous problèmes pour les optimiser chacun individuellement ; et l'évaluation qui consiste à déterminer l'optimum de l'ensemble des solutions réalisables associé au nœud en question ou, au contraire, de prouver mathématiquement que cet ensemble ne contient pas de solution optimale, la méthode la plus générale consiste à borner le coût des solutions contenues dans l'ensemble.[12]

Branch, Cut& Price

Lorsque les variables et les plans de coupes sont générés dynamiquement durant l'algorithme de B&B, on appelle cette technique Branch, Cut & Price. Il existe une certaine symétrie entre la génération de coupes est de variables.[12]

Les méthodes approchées :

Méthodes souvent inspirées de mécanismes d'optimisation rencontrés dans la nature. Elles sont utilisées pour les problèmes où on ne connaît pas d'algorithmes de résolution en temps polynomial et pour lesquels on espère trouver une solution approchée de l'optimum global. Elles cherchent à produire une solution de meilleure qualité possible dictée par des heuristiques avec un temps de calcul raisonnable en examinant seulement une partie de l'espace de recherche.

Dans ce cas l'optimalité de la solution n'est pas garanti ni l'écart avec la valeur optimal. Parmi ces heuristiques, on trouve les métaheuristiques qui fournissent des schémas de résolution généraux permettant de les appliquer potentiellement à tous les problèmes.

Plusieurs classification des métaheuristiques ont été proposées, la plupart distinguent globalement deux catégories : celles se basant sur une solution unique et celles se basant sur une population de solution.

Heuristique

L'heuristique est une méthode, conçue pour un problème d'optimisation donné, qui produit une solution non nécessairement optimale lorsqu'on lui fournit une instance de ce problème.[13]

Méta-heuristique :

Plusieurs définitions ont été proposées pour expliquer clairement la notion de métaheuristique, nous citons parmi elles :

« Un processus itératif qui subordonne et qui guide une heuristique, en combinant intelligemment plusieurs concepts pour explorer et exploiter tout l'espace de recherche. Des stratégies d'apprentissage sont utilisées pour structurer l'information afin de trouver efficacement des solutions optimales, ou presque optimales » [33].

« Les métaheuristiques sont généralement des algorithmes stochastique itératifs, qui progressent vers un optimum global, c'est-à-dire l'extremum global d'une fonction objectif » [34].

Les métaheuristiques constituent une classe de méthodes qui fournissent des solutions de bonne qualité en un temps raisonnable à des problèmes combinatoires réputés difficiles pour lesquels on ne connait pas de méthode classique plus efficace. Elles sont généralement des algorithmes stochastiques itératifs, qui progressent vers un optimum global, c'est à dire l'extremum global d'une fonction en évaluant une certaine fonction objectif. Elles se comportent comme des algorithmes de recherche, tentant d'apprendre les caractéristiques d'un problème à fin d'en trouver une approximation de la meilleure solution d'une manière proche des algorithmes d'approximations.

L'intérêt croissant apporté aux métaheuristiques est tout à fait justifié par le développement des machines avec des capacités calculatoires énormes, ce qui a permis de concevoir des métaheuristiques de plus en plus qui ont fait preuve d'une certaine efficacité lors de la résolution de plusieurs problèmes à caractère NP-difficile.[13]

- **Recherche avec tabous :**

L'idée de cette méthode est de permettre des modifications qui n'améliorent pas la valeur de l'objectif, Toutefois, on choisira toujours la meilleure modification possible, Mais nous venons de voir qu'une des modifications possibles nous ramène à la solution précédente. Il faut donc changer la définition de l'ensemble des modifications possibles pour interdire celles qui nous ramènent à la solution précédente ;

À cette fin, on conservera une liste des dernières modifications effectuées en rendant taboue (en interdisant) la modification inverse, Cette liste taboue peut être vue comme une mémoire à court terme permettant de guider la recherche, A chaque itération, on choisit la meilleure modification possible (excluant celles qui sont taboues), puis on met à jour cette liste en ajoutant la modification inverse de celle effectué ;

Contrairement à la méthode de descente, il n'y a pas de critère d'arrêt simple Typiquement, on utilise une combinaison des critères suivants :

- Nombre maximum d'itérations.
- Temps limite.
- Nombre d'itérations successives sans amélioration.
- Il n'y a plus de modification possible.

Adaptation de cette métaheuristique pour résoudre un problème particulier : structure de voisinage + implantation de la liste taboue.[14]

Recuit simulé « Simulated annealing » :

Comme dans la recherche avec tabous, on permet des modifications qui n'améliorent pas la valeur de l'objectif, Au lieu de choisir la modification la plus intéressante, parmi toutes les modifications possibles, on en choisit une au hasard, On va biaiser le choix vers des modifications qui améliorent ou tout au moins ne détériorent pas tropla valeur de l'objectif.
A partir d'une solution courante, on effectue une modification au hasard qui nous amène à une solution candidate.[14]

Métaheuristiques à base de population de solutions

Travaillent sur un ensemble de points de l'espace de recherche en commençant avec une population de solution initiale puis de l'améliorer au fur et à mesure des itérations. L'intérêt de ces méthodes estd'explorer un très vaste espace de recherche et d'utiliser la population comme facteur «de diversité» .[14]

Algorithmes Evolutionnaires :

Un algorithme évolutionnaire est typiquement composé de trois éléments fondamentaux :

- une population constituée de plusieurs individus représentant des solutions potentielles (configurations) du problème donné, permettant de mémoriser les résultats à chaque étape du Processus de recherche.

- un mécanisme d'évaluation (fitness) des individus permettant de mesurer la qualité de l'individu,

- un mécanisme d'évolution de la population permettant, grâce à des opérateurs prédéfinis(tels que la sélection, la mutation et le croisement), d'éliminer certains individus et d'en créer de nouveaux. Ces méthodes sont applicable dans la plupart des

problèmes d'optimisation (multimodaux, non continu, contraints, bruités, multiobjectif, dynamiques, etc.). [12]

Les algorithmes génétiques :

Appartiennent à la famille des algorithmes évolutionnistes. Leur but est d'obtenir une solution approchée à un problème d'optimisation, lorsqu'il n'existe pas de méthode exacte (ou que la solution est inconnue) pour le résoudre en un temps raisonnable. Les algorithmes génétiques utilisent la notion de sélection naturelle et l'appliquent à une population de solutions potentielles au problème donné. La solution est approchée par « bonds » successifs, comme dans une procédure de séparation et évaluation, à ceci près que ce sont des formules qui sont recherchées et non plus directement des valeurs.**[15]**

Essaim de particulaire PSO :

Met en jeu des groupes de particules sous forme de vecteurs se déplaçant dans L'espace de recherche. Chaque particule p est caractérisée par deux variables d'état (sa position courante x (t)et sa vitesse courante v (t)). Cette technique repose sur deux règles :

- Chaque particule se souvient du meilleur point par lequel elle est passée au cours de ses évolutions et tend à y retourner,
- Chaque particule est informée du meilleur point connu au sein de la population et tend às'y rendre. [12]

-Recherche de coucou (CS)

Est un algorithme d'optimisation, Il a été inspiré par le parasitisme obligatoire de certains coucou espèces par pondre leurs œufs dans les nids d'autres oiseaux d'accueil (d'autres espèces).Certains oiseaux hôtes peuvent engager conflit direct avec les coucous intrus. Certaines espèces de coucou comme le Nouveau Monde couvain-parasitaire Tapera ont évolué d'une manière telle que coucous parasites femmes sont souvent très spécialisé dans le mimétisme dans les couleurs et le modèle des œufs de quelques espèces hôtes choisies, recherche Coucou idéalisé tel comportement de reproduction, et peut donc être appliquée dans divers problèmes d'optimisation. Il semble qu'il peut surpasser les autres algorithmes métaheuristiques dans les applications.

Recherche de coucou (CS) utilise les représentations suivantes :

Chaque œuf dans un nid représente une solution, et un œuf de coucou représente une nouvelle solution. Le but est d'utiliser les nouvelles et potentiellement de meilleures solutions (coucous) pour remplacer une pas-si-bonne solution dans les nids. Dans la forme la plus simple, chaque nid a un œuf. L'algorithme peut être étendu à des cas plus compliqués dans lequel chaque nid a plusieurs œufs représentant un ensemble de solutions.

CS est basé sur trois règles idéalisées :

1. Chaque coucou pond un œuf à la fois, et le décharge dans un nid choisi au hasard ;
2. Les meilleurs nids de haute qualité des œufs seront reportés à la prochaine génération ;
3. Le nombre d'hôtes disponibles est fixé, et l'œuf pondu par un coucou est découvert par l'oiseau hôte avec une probabilité $p_a \in (0,1)$.

IV. Conclusion :

Dans ce chapitre, nous avons présenté deux parties, la première partie aborde la notion d'apprentissage et les algorithmes associés et dans la deuxième partie, nous avons précisé généralement la notion de base de l'optimisation puis l'optimisation combinatoire avec la complexité des problèmes et les différentes méthodes utilisées pour résoudre ces problèmes.

Chapitre II :

La bioinformatique et la sélection d'attributs

I. Introduction :

Au cours de ces trente dernières années, la récolte de données en biologie a connu un boom quantitatif notamment grâce au développement de nouveaux moyens techniques servant à comprendre l'ADN et d'autres composants d'organismes vivants. Pour analyser ces données, plus nombreuses et plus complexes aussi, les scientifiques se sont tournés vers les nouvelles technologies de l'information. L'immense capacité de stockage et d'analyse des données qu'offre l'informatique leur a permis de gagner en puissance pour leurs recherches. La rencontre entre la biologie et l'informatique, c'est ce qu'on appelle la bioinformatique.

La sélection d'attributs est l'une des techniques importantes et fréquentes utilisée dans le prétraitement de données.

Dans ce chapitre, nous allons présenter quelques notions de base de la bioinformatique, ensuite, nous exposerons ce que la découverte et les étapes du développement des bio marqueurs, et à la fin en termine par le processus de la sélection d'attributs.

II. La bioinformatique :

1. La bioinformatique, c'est quoi ?

La bioinformatique est l'analyse de la bioinformation.

La bioinformation est l'information liée aux molécules biologiques : leurs structures, leurs fonctions, leurs liens de "parenté", leurs interactions et leur intégration dans la cellule.

Divers domaines d'études permettent d'obtenir cette bioinformation : la génomique structurale, la génomique fonctionnelle, la protéomique, la détermination de la structure spatiale des molécules biologiques, la modélisation moléculaire ...etc .[16]

2. Comment ça marche ?

"La bioinformatique fournit des bases de données centrales, accessibles mondialement, qui permettent aux scientifiques de présenter, rechercher et analyser de l'information. Elle propose des logiciels d'analyse de données pour les études de données et les comparaisons et fournit des outils pour la modélisation, la visualisation, l'exploration et l'interprétation des données", selon une définition de l'Institut Suisse de Bioinformatique.[17]

3. Ça sert à quoi ?

La bioinformatique sert donc à stocker, traiter et analyser de grandes quantités de données de biologie. Le but est de mieux comprendre et mieux connaître les phénomènes et processus biologiques. Grâce à ces nouvelles connaissances ainsi acquises, les chercheurs ont la possibilité de faire de nouvelles découvertes scientifiques. Des découvertes qui peuvent par exemple améliorer la qualité de vie de personnes malades grâce à la mise en place de nouveaux traitements médicaux plus efficaces.[17]

4. Les molécules supportées par la bioinformation :

Deux types de molécules support de la bioinformation : les acides nucléiques (ADN et ARN) et les protéines :

ADN : Acide Désoxyribonucléique	**ARN : Acide Ribonucléique**	**Protéine**
-macromolécule : chaîne nucléotidique -constituée par un enchaînement d'unités élémentaires : les désoxyribonucléotides - forme de stockage de l'information génétique. Cette information est représentée par une suite linéaire de gènes. - formée de deux brins complémentaires enroulés en double hélice ce qui lui permet de se dupliquer en deux molécules identiques entre elles et identiques à la molécule mère. - On distingue : • l'ADN du génome du noyau • l'ADN du génome mitochondrial • l'ADN du génome chloroplastique	-macromolécule : chaîne nucléotidique -constitué par un enchaînement d'unités élémentaires : les ribo nucléotides -forme qui permet de transférer et de traiter l'information dans la cellule -le plus souvent formé d'un simple brin -On distingue : • les ARN messagers : ils sont transcrits à partir d'un gène (ADN). Ils sont ensuite traduit en protéine les ARN de transfert • les ARN ribosomaux • les ARN nucléaires	-macromolécule : chaîne polypeptidique -constitué par un enchaînement d'unités élémentaires : les acides aminés. -l'ensemble des protéines assurent les principales fonctions cellulaires -se replie sur elle-même et adopte une conformation ou structure particulière dans l'espace. Cette structure tridimensionnelle est à l'origine de la fonction de la protéine et de la spécificité de cette fonction.

	• Les ARN cytoplasmiques	

Tableau 1 :Les molécules support par la bioinformation. [16]

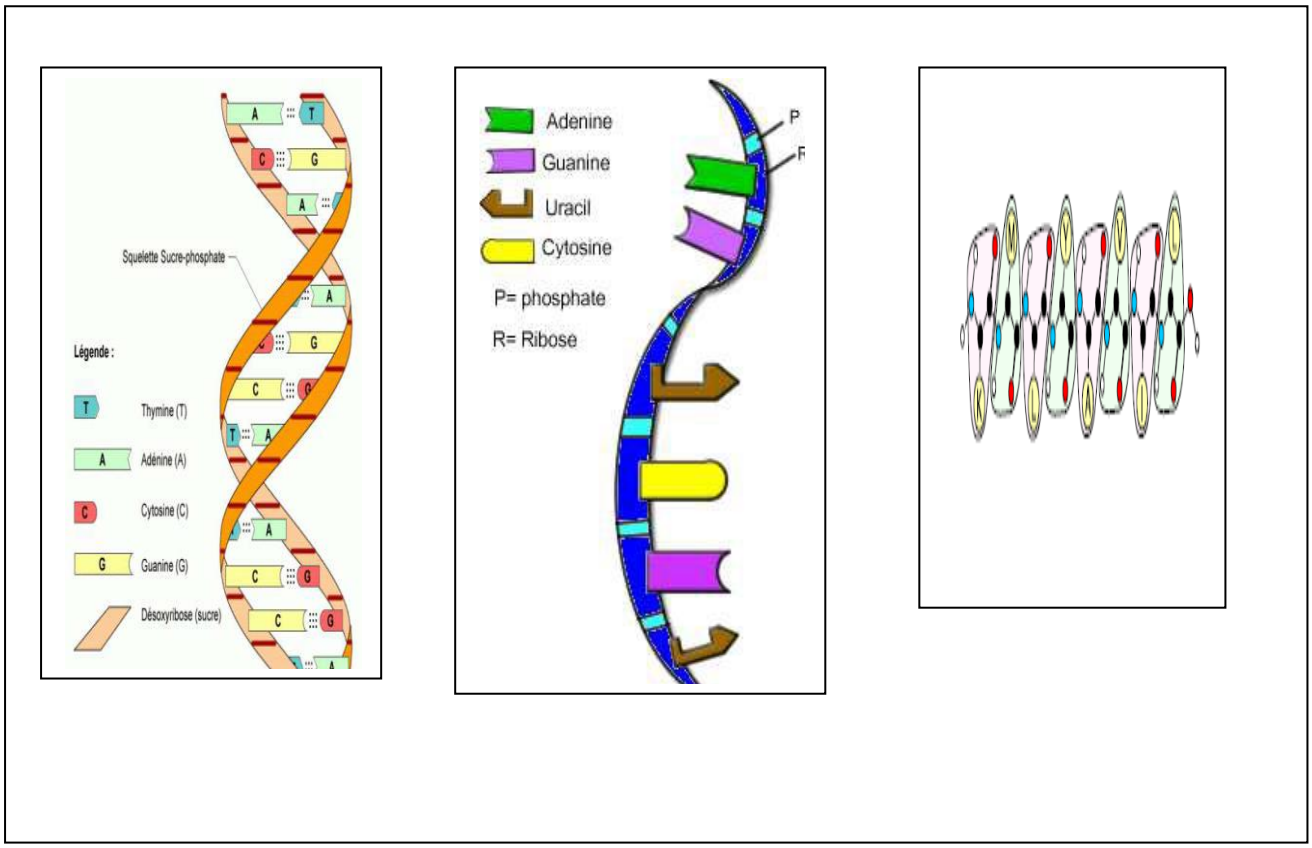

Figure 8 : la structure des ADN, ARN et Protéine

5. Les banques et les bases de données biologiques :

Les bases de données ont pour mission de rendre publiques la masse des données qui ont été déterminées. Par conséquent la quantité de données annotées qu'elles contiennent, le nombre de séquences, dont certaines ne sont plus disponibles ailleurs, leur grande diversité et la qualité des annotations en font des outils indispensables pour la communauté scientifique. Et leur utilisation se trouve au cœur des pratiques en matière de recherche. Voilà pourquoi leur utilisation est nécessaire au progrès scientifique. En biologie on distingue généralement deux types de bases de données biologiques :[18]

Les Bases généralistes :

Ce sont des Bases publiques internationales qui recensent des séquences d'ADN ou de protéines déterminées quelque soit l'organisme ou la méthode utilisée pour les acquérir. Elles présentent cependant quelques défauts qui ont conduit à la création de bases plus spécifiques : Les données qu'elles contiennent n'ont pas toujours été vérifiées, elles sont parfois trop diverses et nombreuses pour être exploitées efficacement et sont parfois mises à jour avec un certain retard.[18]

Les Bases spécialisées :

Devant la croissance quasi exponentielle des données et l'hétérogénéité des séquences contenues dans les principales bases de séquences généralistes, d'autres bases spécialisées sont apparues. Ce sont des bases dont les données ont été classées suivant une caractéristique biologique particulière comme les signaux de régulation, les régions promoteurs de gènes, les signatures peptidiques ou les gènes identiques issus d'espèces différentes.
L'utilisation des bases spécialisées comme les bases de motifs est devenue un outil essentiel dans l'analyse des séquences pour tenter de déterminer la fonction de protéines inconnues ou savoir à quelle famille appartient une séquence non encore caractérisée.[18]

6. La structuration de la bioinformation : fichiers et formats

Les séquences sont stockées en général sous forme de fichiers texte qui peuvent être soit des fichiers personnels (présents dans un espace personnel), soit des fichiers publics (séquences des banques) accessibles par des programmes interfaces (tels que SRS, GCG).

Le format correspond à l'ensemble des règles (contraintes) de présentation auxquelles sont soumises la ou les séquences dans un fichier donné.

Le format permet :

- Une mise en forme automatisée
- Le stockage homogène de l'information
- Le traitement informatique ultérieur de l'information.
- Pour lire et traiter les séquences, les logiciels d'analyse autorisent un ou plusieurs format des données.[16]

7. applications actuelles de la bioinformatique :

On peut classer ces applications en différentes catégories compte-tenu de la diversité des domaines d'action de la bioinformatique.

- L'une des premières applications est bien sûr l'**analyse de séquences** qui peut aller de l'identification de gènes aux comparaisons de séquences en passant par la prédiction de motifs ou l'établissement de signatures.
- La **structure des protéines** nécessite l'usage de la bioinformatique, que ce soit pour la visualisation ou la prédiction de leur repliement.
- Un autre aspect de la bioinformatique réside dans son utilisation en **phylogénie** de façon à comparer les espèces à l'échelle moléculaire et obtenir ainsi un classement évolutif qui soit plus fiable.
- Des **liaisons génétiques** peuvent être établies grâce à la bio-informatique pour permettre de détecter des gènes candidats de maladies génétiques par exemple.
- Enfin, l'utilisation de la bio-informatique en **génomique fonctionnelle** permet de travailler sur le transcriptome, le protéome, l'intéractome...etc. [19]

III. Découverte de biomarqueurs

La **découverte de biomarqueurs** est un terme médical qui décrit le processus par lequel les biomarqueurs sont découverts. Il y a intérêt pour la découverte de biomarqueurs de la part de l'industrie pharmaceutique ; test sanguin ou d'autres biomarqueurs pourraient servir de marqueurs intermédiaires de la maladie dans les essais cliniques, et comme possibles cibles médicamenteuses.

1. Définition du biomarqueur :

Un biomarqueur est une caractéristique mesurable objectivement qui représente un indicateur des processus biologiques normaux ou pathologiques ou de réponse pharmacologique à une intervention thérapeutique. [20]

2. Types de biomarqueurs :

Les biomarqueurs sont utilisés en cancérologie pour [35]:

1. Aider au diagnostic, par exemple à l'identification des cancers dans les premiers stades de leur développement
2. Prévoir l'agressivité de la tumeur, de même que la chance de s'en sortir en absence de traitement (pronostic)
3. Prédire la réponse du patient au traitement proposé (prédiction)
4. Définir la dose optimale de médicaments chez un patient donné (pharmaco-dynamique prédictible)
5. Déterminer le risque de la récurrence (pronostic). [20]

3. Les différentes étapes du développement des biomarqueurs (moléculaires) :

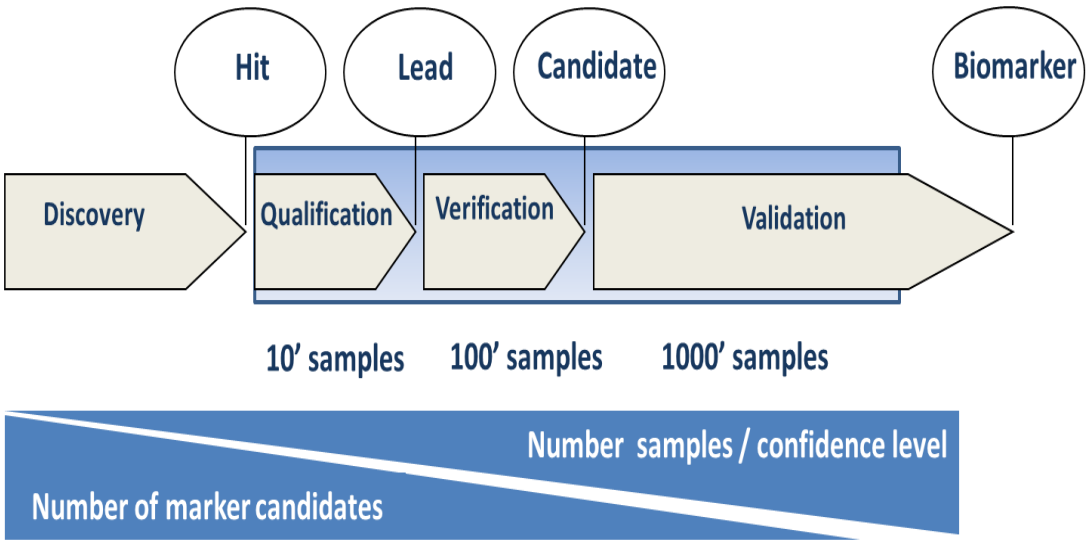

Figure 9 : les étapes du développement des biomarqueurs

Découverte

-Détection et identification des différentes molécules exprimées dans plusieurs populations.

-Etudes rétrospectives requises sur plusieurs dizaines d'échantillons.

-Conclusion de la phase de découverte : un "hit".

Qualification

-Confirmation de l'expression différentielle des molécules dans un même échantillon en utilisant une technique différente.

-Etudes rétrospectives requises sur plusieurs dizaines d'échantillons.

-Conclusion de la phase de qualification : un "lead".

Vérification :

-Vérification et quantification statistique du différentiel d'expression des molécules "marqueuses" sur des cohortes de plusieurs centaines de sujets.

-Etudes rétrospectives requises sur des centaines d'échantillons.

-Résultat de la phase de vérification : biomarqueur candidat.

Validation :

-Preuve statistique et quantification du test candidat sur des cohortes de mille (et plus) sujets sains atteints d'une pathologie spécifique, atteints d'une pathologie similaire...

-Etudes préférentiellement prospectives requises sur des milliers d'échantillons.

-Résultat de la phase de validation : Biomarqueur cliniquement validé.[20]

IV. la sélection d'attributs :

1. Quelques définitions :

Dash propose de regrouper les techniques de sélection d'attributs en fonction de l'objectif visé.

Il identifie alors quatre classes distinctes :

- « Classic » : La sélection d'attributs est un processus qui permet de réduire l'ensemble des attributs de N à M tel que M < N, et la fonction d'évaluation soit optimal. Cette réduction de la dimensionnalité mène à une accélération de la vitesse du traitement et augmentation et amélioration de la précision du traitement.
- *«Idealized* » : La sélection d'attributs est un processus qui permet de trouver le sous ensemble de taille minimale qui est nécessaire et suffisant pour atteindre l'objectif fixé.

- « *Improving prediction accuracy* » : La sélection d'attributs est un processus qui permet de choisir un sous-ensemble d'attributs afin d'améliorer la précision de la prédiction ou diminuer la taille de la structure sans diminution significative de la précision de prédiction du classificateur, construit en utilisant seulement les variables sélectionnées.
- « *Approximating original class distribution* » : La sélection d'attributs est un processus qui permet de sélectionner un sous-ensemble de variables tel que la distribution des classes résultante soit aussi proche que possible de la distribution des classes étant donné l'ensemble des variables complet.[24]

2. Pourquoi la sélection d'attributs :

On utilise la sélection d'attributs pour des facteurs sans influence ou peu influents et les facteurs redondants en plus en utilise la sélection d'attributs pour la dimension des entrées telle que coût de l'apprentissage trop grand donc apprentissage moins coûteux et aussi la sélection d'attributs joue un rôle pour faciliter l'apprentissage tel que elle offre des meilleures performances en classification, meilleure compréhensibilité de l'hypothèse ,et identification des facteurs pertinents (Génomique Vision).[22]

4. Le cadre général d'un algorithme de sélection d'attributs :

Les étapes de la sélection d'attributs sont :

Etape1 : génération de sous ensemble d'attribut.

Etape2 : Stratégie d'évaluation.

Etape3 : Critère d'arrêt.

Etape4 : Validation des résultats

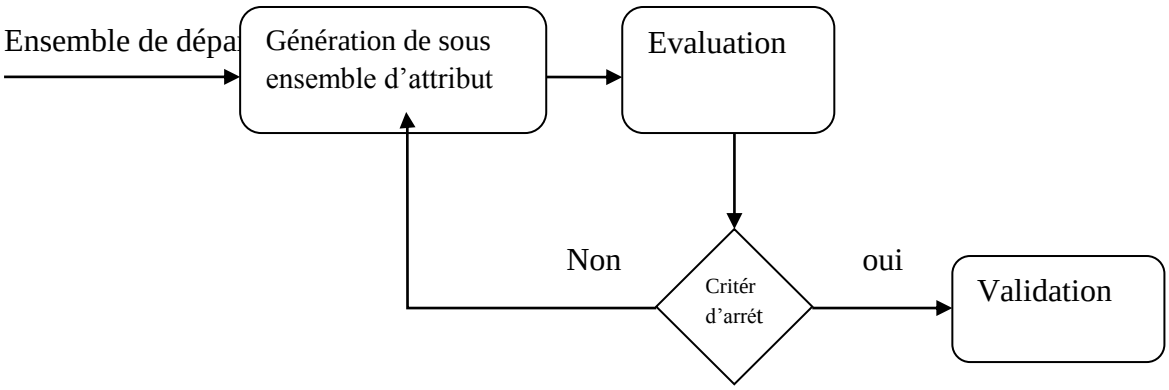

Figure 10 : Processus de la sélection d'attributs[24]

3.1. Génération de sous ensemble :

Est un processus qui produit un sous ensemble d'attributs candidat pour évaluation, la nature de celui-ci est déterminée par les deux étapes suivante :

Point de départ

Est le point dans l'espace des sous-ensembles d'attributs, à partir duquel on commence la recherche, et ce point même qui va affecter la direction de recherche.

Pour N attributs, l'espace de recherche contient 2N-1 sous-ensembles possibles.

-Si on commence avec zéro attribut, on doit faire des rajouts successifs d'attributs et c'est l'*Option forward*.

-Si on commence avec l'ensemble de tous les attributs, on doit faire des suppressions successives et c'est *l'Option backward.*

-Si on commence avec quelques attributs on doit faire des rajouts et des suppressions et c'est l'*Option stepwise.*

Une stratégie de recherche :

Est une procédure qui permet d'explorer l'espace des combinaisons des attributs.

Cet espace de recherche est exponentiellement prohibitif pour la recherche exhaustive.

Pour cela différentes stratégies de recherche sont à explorer, on peut les classifier en trois catégories :

- **La recherche complète :**

Elle garantit le résultat optimal par rapport au critère d'évaluation utilisé. Une recherche exhaustive est complète mais la recherche ne doit pas être exhaustive pour qu'elle soit complète. Différents heuristiques et fonctions peuvent être utilisées pour réduire l'espace de recherche sans avoir le risque de perdre les résultats optimaux

- **La recherche séquentielle :**

Elle ne garantit pas le résultat optimal, elle consiste à rajouter ou éliminer itérativement des attributs.

Il existe plusieurs variations de l'approche greedyhill-climbing comme :

***Forward*:** cette approche part d'un ensemble d'attributs vide auquel, à chaque itération sont ajoutés un ou plusieurs attributs. Elle est également appelée approche ascendante[24]

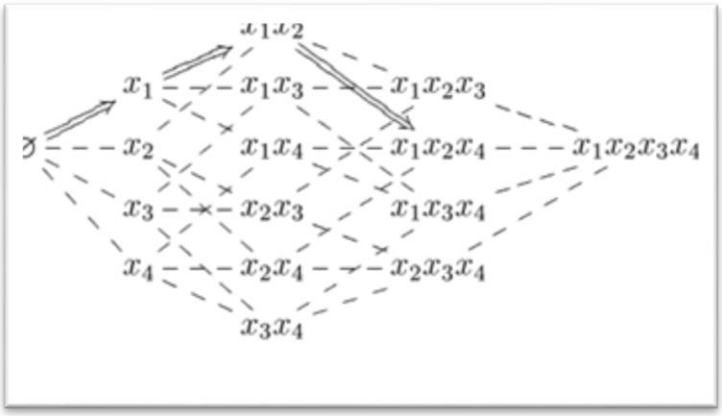

Figure 11: Sélection d'attributs Forward.

***Backward*:** c'est l'approche inverse ; l'ensemble total des attributs est considéré au départ de la procédure itérative, chaque itération permet d'en supprimer. Une autre appellation est approche descendante [24]

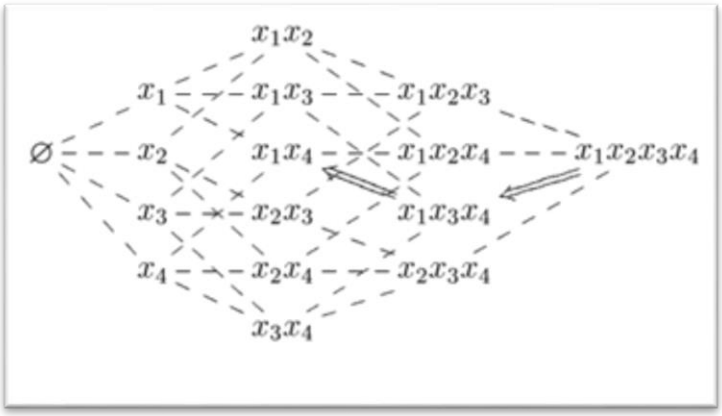

Figure 12: Sélection d'attributs Backrward.

Bidirectionnelle :

Les méthodes SFS et SBS sont connus par leur simplicité de mise en œuvre et leur rapidité. Cependant, comme elles n'explorent pas tous les sous-ensembles possibles d'attributs et ne permettent pas de retour arrière pendant la recherche ; elles sont donc sous-optimales.

Pour réduire cet effet, il y a d'autres alternatives qui permettent d'ajouter des attributs et d'en retirer d'autre.

Les méthodes flottantes SFFS et SFBS sont une extension de l'algorithme "*plus l-takeawayr*". Elles sont considérées comme les méthodes sous-optimales les plus efficaces.

- ***SFFS:*** c'est l'acronyme de sequential forward floating search. L'algorithme SFFS consiste à appliquer après chaque étape *forward* autant d'étapes *backward* que le sous-ensemble d'attributs *F* correspondant améliore le critère d'évaluation *J*(*F*) à ce niveau de recherche.
- ***SBFS :*****c'**est l'acronyme de sequential forward floating search. Le même principe de SFFS est appliqué sauf que les deux étapes sont inversées.[24]

La recherche aléatoire :

Est la plus récente dans le domaine de la sélection d'attributs.

Chaque sous ensemble d'attributs est généré d'une manière complètement aléatoire.

L'obtention de bons résultats à l'aide de ces techniques nécessite un choix judicieux de ces paramètres.

Comparaison entre méthodes de recherches :

	Rigueur (Exactitude)	**Complexité**	**Avantages**	**Inconvénients**
Exhaustive	Trouve toujours la solution optimale	Exponentielle	Grande Exactitude	Complexité très Elevé
Séquentielle	Bonne s'il n'y a pas de retour arrière	Quadratique O(NEX2)	Simple et rapide	Pas de retour Arrière
aléatoire	Bonne avec un bon contrôle de paramètres	Généralement basse	Désigné pour échapper des optima locaux	Difficile de régler les paramètres

Tableau 2 : comparaison entre méthodes de recherches[24]

3.2 L'évaluation du sous ensemble :

L'amélioration des performances d'un système d'apprentissage par une procédure de sélection de variables nécessite dans un premier temps la définition d'une mesure de pertinence ou bien un critère d'évaluation.

Typiquement, une fonction d'évaluation essaie de mesurer le pouvoir discriminant d'une variable ou d'un ensemble de variables pour discerner entre les différentes classes.

Par contre, pour un problème de régression, on teste plutôt la qualité de prédiction par rapport aux autres variables. La pertinence d'une variable (ou d'un ensemble de variables) peut être définie par :

Définition de pertinence d'une variable : Une variable pertinente est une variable telle que sa suppression entraîne une détérioration des performances du pouvoir de discrimination en classement ou la qualité de prédiction en régression du système d'apprentissage.

Toutefois, le choix d'un sous-ensemble optimal résultat de la procédure de recherche est relatif à la fonction d'évaluation utilisée. Ainsi, le changement du critère peut changer l'ensemble optimal en résultat.

Dès lors, plusieurs critères d'évaluation basés sur des hypothèses statistiques ou sur des heuristiques ont été proposés. Dans le cadre d'un problème de classification, les critères d'évaluation sont souvent basés sur les matrices de dispersion intra et inter classes, qui sont liées à la géométrie et la distribution des classes dans l'espace.

D'autres critères d'évaluation utilisent des distances probabilistes ou des mesures d'entropie basées sur l'information mutuelle entre les variables et les classes des observations [Slonim and Tishby, 1999].

Les méthodes de sélection peuvent être classées en deux grandes approches (les filtres et les wrappers), selon leur dépendance vis-à-vis l'algorithme inductif qui utilisera par la suite le sous-ensemble optimal des attributs [Kohavi and John, 1997].

Les méthodes filtres sont indépendantes de l'algorithme inductif, alors que les méthodes wrappers utilisent l'algorithme inductif comme une fonction d'évaluation.

Selon, les fonctions d'évaluation peuvent être divisées en 5 catégories :

- **Mesure de distance :** c'est une mesure de séparabilité, divergence ou bien mesure de discrimination comme par exemple la distance Euclidienne. Dans le cas d'une classification binaire, un attribut X est préféré à un autre attribut Y si X induit une plus grande différence entre la probabilité conditionnelle des deux classes en question.
- **Mesure d'information** : cette mesure détermine l'information apportée par un attribut comme par exemple la mesure d'entropie. L'information apportée par un attribut X est déterminée comme étant la différence entre l'incertitude préalable et l'incertitude postérieure en utilisant X. Ainsi, un attribut X n'est préféré à un autre attribut Y que si l'information apportée par X est plus que celle apporté par Y.
- **Mesure de dépendance :** c'est la mesure de corrélation qui peut qualifier la capacité de prédire la valeur d'une variable depuis une autre variable. Si la corrélation entre un attribut

X et une classe C est supérieure à celle entre un attribut Y et la classe C, alors X est préféré à Y.

- **Mesure de consistance :** il est proportionnel au pouvoir discriminant. Un sous-ensemble de variables ayant un taux d'inconsistance élevé signifie que ces variables ne permettent pas de bien prédire la classe et donc que ce sous-ensemble n'est pas un bon ensemble discriminant [36].

- **Mesure d'erreur de classification :** ce sont les méthodes wrappers qui utilisent ce type de fonction d'évaluation. Les attributs permettant d'améliorer l'erreur de classification sont sélectionnés. Ainsi, une grande précision de classification est garantie mais en revanche d'un calcul couteux en temps et mémoire [23]

3.3. Procédure de validation :

Une manière de faire la validation des résultats est de mesurer directement ces derniers en utilisant des connaissances a priori sur les données. Si nous connaissons l'attribut pertinent à l'avance comme dans le cas des données synthétiques, nous pouvons comparer cet ensemble d'attributs connu avec les attributs sélectionnés (Les connaissances sur les attributs non pertinents ou redondants peuvent également aider). Dans les applications du monde réel, habituellement nous n'avons pas de connaissances a priori. Par conséquent, nous devons compter sur quelques méthodes indirectes en surveillant le changement des performances par rapport les changements des attributs. Par exemple, si nous employons le taux d'erreur de classification comme un indicateur de performance pour le traitement, pour un sous-ensemble d'attributs sélectionné, nous pouvons simplement suivre l'expérience "avant et après" pour comparer le taux d'erreur de classificateur sur l'ensemble complet d'attributs et sur le sous ensemble sélectionné.[21]

3.4. Condition d'arrêt :

Le nombre optimal de variables à sélectionner n'est pas connu a priori. Ce qui fait que l'utilisation d'une règle pour contrôler la sélection de variables (sous-ensemble de variables) permettra d'arrêter la recherche dans le cas où aucune variable (sous-ensemble de variables) n'est plus suffisamment informative.

Le critère d'arrêt est souvent défini en fonction d'une combinaison de la procédure de recherche et du critère d'évaluation. Prédéfinir un nombre maximal d'itérations à ne pas

franchir est un critère d'arrêt assez commun. Cependant ce critère peut arrêter la recherche trop tôt ou bien à l'inverse trop tard. Un nombre maximal d'attributs peut aussi être utilisé comme critère d'arrêt. Cependant l'estimation du nombre optimal d'attributs n'est pas donnée préalablement. Définir le critère à base de la fonction d'évaluation est aussi envisageable. Dans ce cas, un seuil est fixé préalablement pour contrôler la variation de la fonction d'évaluation entre deux itérations consécutives.[23]

4. Cadre de catégorisation :

[37] ont développé un cadre de catégorisation en trois dimensions.

Les critères d'évaluation et les stratégies de recherche sont des facteurs dominants pour la conception des algorithmes de sélection d'attributs (certains exemples d'algorithme sont présentés en annexe B)

- Suivant les stratégies de recherches : les algorithmes sont classés par catégories en complet, séquentiel et aléatoire.
- Sous les critères d'évaluation : les algorithmes sont classés par catégories Filter, Wrapper et Hybride.
- La troisième dimension est les algorithmes d'induction car la disponibilité de l'information de classe dans la classification ou clustering affecte un critère d'évaluation utilisé dans les algorithmes de sélection des attributs.

Avec la catégorie Filter, on distingue des critères d'évaluation spécifique : distance, information, dépendance et cohérence.

Avec la catégorie Wrapper, il y a la prédiction de la précision (utilisée pour la classification) et cluster goodness (utilisé pour le clustering).

Trois rôles présentés pour ce cadre :

1. Trouver les relations entre les algorithmes.
2. Choisir un algorithme de sélection d'attributs pour une tâche donnée parmi ces algorithmes.
3. Trouver les combinaisons inexplorées des procédures de génération et d'évaluation.

Les méthodes Filtrantes

Ces méthodes sélectionnent les attributs en utilisant les différentes approches et les différents critères pour calculer la pertinence d'un attribut avant le processus d'apprentissage c'est-à-dire la construction d'un classifieur.

Les méthodes hybrides

Elles sont récemment proposées pour une large base de données **[38].** L'algorithme hybride commence la recherche à partir d'un sous ensemble S0 dans la sélection séquentielle «

Forward ». Cet algorithme fait l'itération pour trouver le meilleur sous ensemble à chaque augmentation de cardinalités.

Les méthodes Wrapper

Ces méthodes se servent de l'algorithme d'induction comme d'une boite noire : l'apprentissage est effectué avec les variables sélectionnées et les performances sont estimées à partir de l'erreur de généralisation. La méthode Wrapper conduit une recherche dans l'espace des paramètres possibles.

Une recherche requiert **:**

- un espace d'états où chaque état représente un sous ensemble d'attributs. Pour n attributs, il y a n bits dans chaque état et chaque bit indique si l'attribut est présent ou absent.
- un état initial : lorsque l'on fait une sélection Forward, la recherche commence avec un ensemble vide d'attributs, lorsqu'on fait une élimination « backward », la recherche commence avec l'ensemble complet d'attributs.
- une condition d'arrêt.
- une méthode de recherche.

Algorithme Wrapper [37] utilise l'algorithme d'induction au lieu de la mesure d'indépendance M pour l'évaluation de sous ensemble S. Il évalue chaque sous ensemble généré S par sa qualité utilisant l'algorithme d'induction sur les données et évalue la qualité des résultats

1) L'approche de type **Wrapper** utilise le classifieur pour évaluer le sous ensemble d'attributs choisis

Les avantages : la méthode Wrapper peut être utilisée lorsqu'on travaille avec un très grand nombre d'attributs car elle est de complexité raisonnable. Elle ne tient que des informations présentées dans les données et elle est indépendante du processus de la classification.

Les inconvénients : la méthode Wrapper repose sur le choix d'un seuil pour le critère de pertinence choisi ou d'un nombre d'attributs à choisir, le choix de ces paramètres n'est pas facile à réaliser.

Algorithme Wrapper

Entrées: D(F0, F1, …, Fn-1) //l'ensemble des données d'apprentissage avec N Attributs

S0 // un sous ensemble initial pour commencer la recherche

δ // critère d'arrêt

Sorties:Sbest // un sous ensemble optimal

Début

Initialiser : Sbest= S0 ;

γbest=eval(S0, D, A); //évaluer S0 par un algorithme de recherche A

Répéter

S=générer(D) ; // générer un sous ensemble pour l'évaluation

γ=eval(S, D, A); //évaluer le sous ensemble actuel S par A

si (γ est meilleur que γbest) alors

γbest= γ ;

Sbest=S ;

Jusqu'à (δ est atteinte) ;

Retourner Sbest;

Fin

Algorithme 1 : Algorithme Wrapper [Liu et Yu, 2005]

2)-L'approche de type Filtrante utilise une fonction spécifique pour évaluer le sous ensemble d'attributs choisi

Les avantages : Le sous ensemble choisi est parfaitement adapté au classifieur.

Les inconvénients : Il y a un risque de Sur-Apprentissage. Elle est plus couteuse en temps de calcul (construction du classifieur à chaque évaluation de sous ensemble candidats). La complexité de calcul dépend de la complexité du modèle d'apprentissage utilisé.[13]

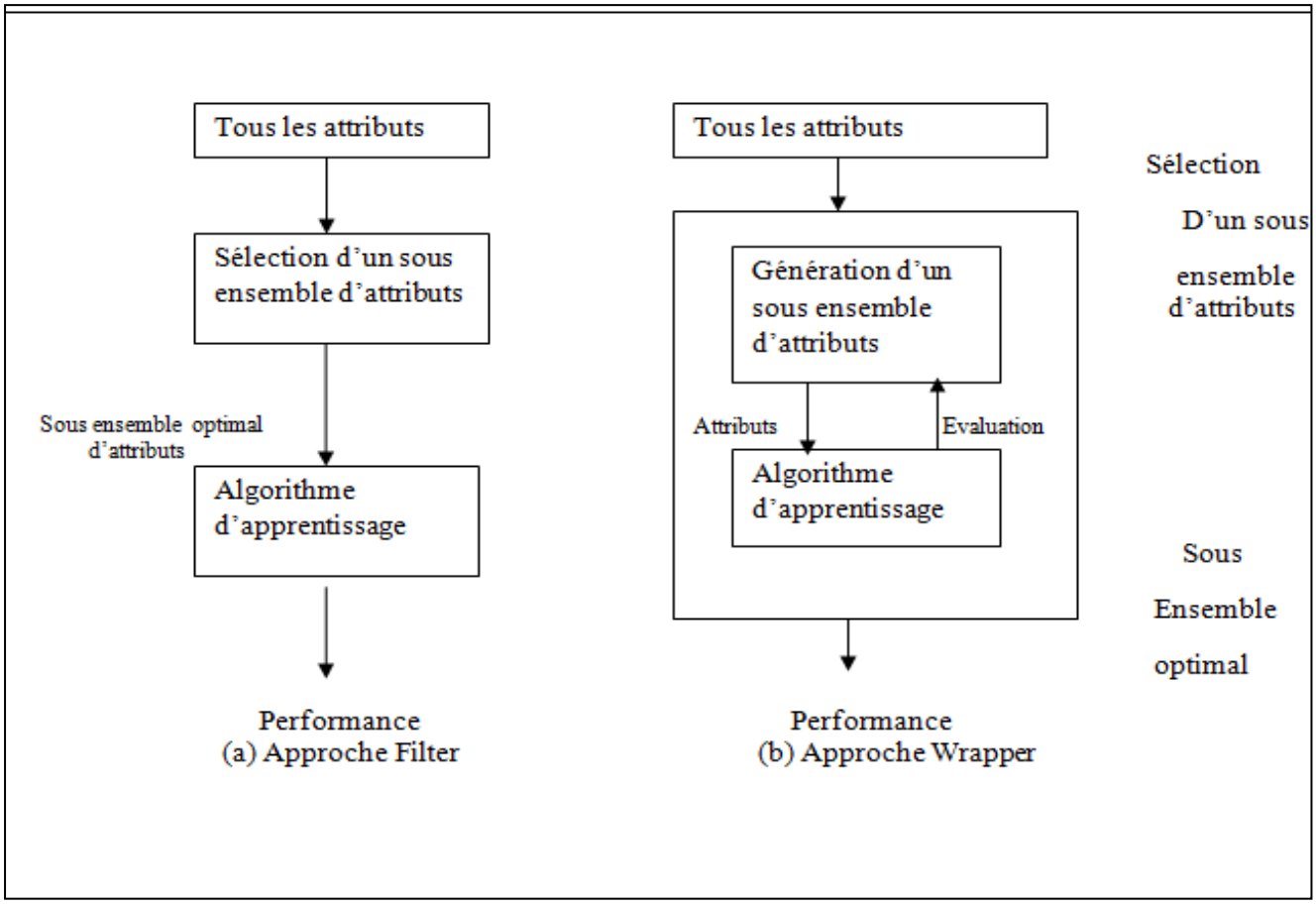

Figure 13 : les deux approches de la sélection d'attributs.[13]

V. Conclusion :

Actuellement dans le domaine de la bioinformatique, la taille des bases de données croît exponentiellement, notamment en raison du séquençage des génômes, il est donc nécessaire d'avoir une technique pour exploitation de ces données.

Nous avons présenté une étape de prétraitement « la sélection d'attributs » qui joue un rôle important. Elle permet de construire un modèle décrivant les données en supprimant les attributs redondants, non pertinents ou bruités.

Chapitre III

L’hybridation de l’Apprentissage et l’Optimisation

I. Introduction :

Le théorème « No free lunch » qui a été introduit en 1997 par David H. Wolpert et William G. Macready [Wolpert et Macready, 1997] stipule qu'aucune méthode n'est systématique (efficace 100%) pour la résolution de tous les problèmes d'optimisation. En effet, chaque méthode a ses avantages et ses inconvénients. Plusieurs communautés de chercheurs ont envisagé la combinaison des méthodes de résolution des problèmes, afin de tirer profit des avantages de chaque méthode et de combler certaines de ses lacunes. Par conséquent, ils ont donné naissance à une nouvelle classe de méthodes de résolution de problèmes d'optimisation : c'est la classe des méthodes hybrides.

Alors, dans notre travail, nous avons choisi la combinaison de deux méthodes, c'est l'optimisation par essaim particulaire (PSO) ; l'algorithme de recherche coucou amélioré (ICS) avec l'algorithme machines à vecteurs de support (SVM) pour la sélection d'attributs.

Dans ce chapitre, notre allons faire une étude générale sur l'algorithme d'optimisation par essaim particulaire (PSO) ; l'algorithme de recherche coucou amélioré cuckoo search (ICS), et nous exposerons l'algorithme machines à vecteurs de support.

II. Notion d'hybridation

L'hybridation consiste à combiner les caractéristiques de deux ou plusieurs méthodes différentes pour en tirer les avantages. Actuellement, les méthodes hybrides sont devenues plus populaires car les meilleurs résultats trouvés pour plusieurs problèmes d'optimisation combinatoire ont étés obtenus avec des algorithmes hybrides.

En effet, Pour définir une méthode hybride efficace, il faut savoir caractériser les avantages et les limites de chaque méthode. Par exemple, les algorithmes génétiques sont très performants lorsqu'il s'agit d'explorer l'espace de recherche, mais ils s'avèrent ensuite incapable d'exploiter efficacement la zone vers laquelle la population des solutions converge.

Il est alors plus intéressant d'utiliser dans ce stade une autre méthode permettant une bonne exploitation comme par exemple le recuit simulé ou une autre heuristique d'amélioration. Il faut souligner qu'il faut être prudent sur le choix des méthodes à hybrider ainsi sur le problème de multiplication des paramètres.

Les approches hybrides ont permis d'obtenir de bons résultats dans une grande variété de problèmes théoriques d'optimisation combinatoire tels le problème du voyageur de commerce, le problème de coloration de graphe, le problème d'affectation quadratique, le problème de tournée de véhicules, le séquençage d'ADN ou encore le calcul des trajectoires des satellites.

On peut citer aussi que l'hybridation de métaheuristiques est la voie la plus prometteuse pour l'amélioration de la qualité des solutions dans beaucoup d'applications réelles. Ainsi, Le choix d'une approche hybride devient aujourd'hui déterminant pour obtenir de meilleures performances lors de la résolution des problèmes complexes.[25]

III. Classification des stratégies d'hybridation :

En général, l'hybridation de deux ou plusieurs techniques pour la résolution d'un problème peut s'effectuer selon trois modes d'association possibles :

1. **L'hybridation séquentielle** :

Les premières techniques génèrent les valeurs d'entrée de la dernière.

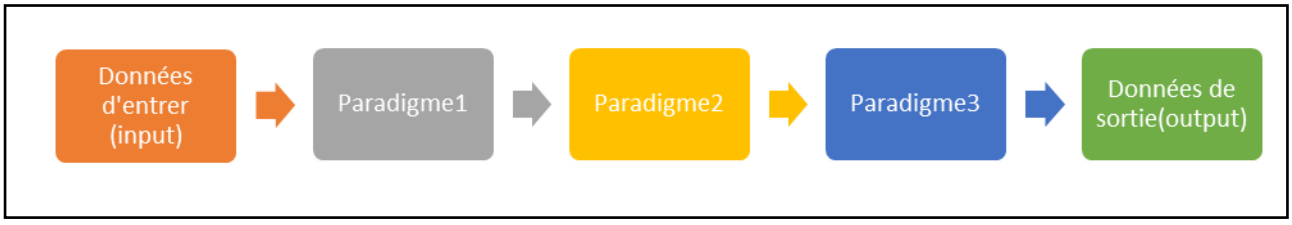

Figure 14 : stratégie de l'hybridation séquentielle [26]

2. L'hybridation auxiliaire :

Une technique sollicite l'aide d'une deuxième ou d'une troisième qui lui renvoie de l'information.[26]

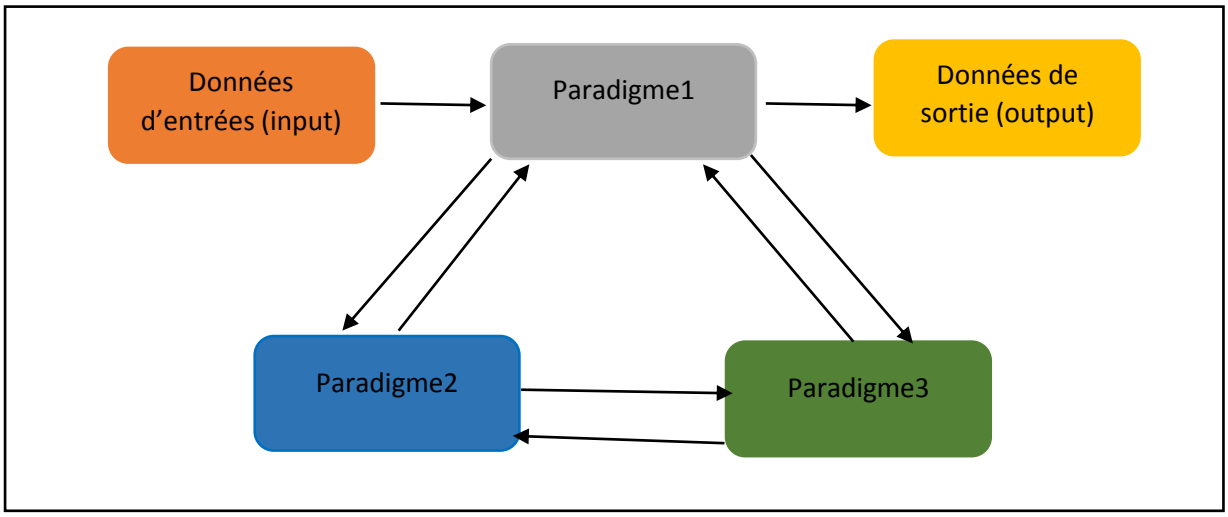

Figure 15 : stratégie de l'hybridation auxiliaire. [26]

Même si le paradigme 1 parvenait à résoudre un problème donné, son association avec un ou deux autres paradigmes peut parfois améliorer de façon significative la procédure et la qualité de sa résolution.

3. L'hybridation emboîtée :

Une technique en englobe d'autres pour former un module plus complexe mais, idéalement, plus fonctionnel.

Puisque les deux ou trois paradigmes s'avèrent essentiels à la résolution du problème, c'est le type d'hybridation la plus complète. [26]

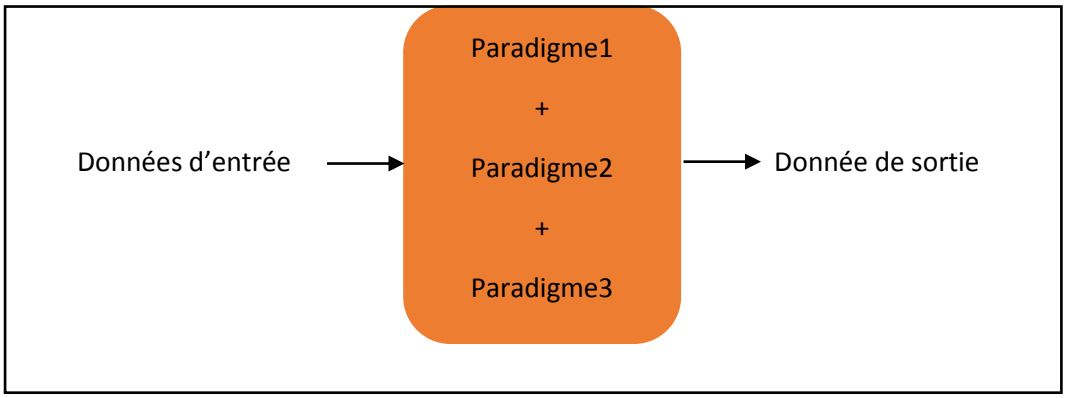

Figure 16 : la stratégie hybridation emboité. [26]

IV. L'hybridation de techniques :

Le tableau suivant permet de se faire une idée assez précise des cas d'hybridations obtenues avec les réseaux de neurones artificiels, la logique floue, les algorithmes génétiques et les ensembles approximatifs :

	Caractéristiques	Commentaires
Systèmes neuro-flous	Prise en charge par la logique floue des pseudo-règles telles développées et exprimées par le réseau de neurones. Ils forment la base des systèmes adaptatifs / intelligents.	Les avantages découlant de la combinaison logique floue/réseaux de neurones permettent de réduire les désavantages propres à chacune de ces technologies prises isolément.
	Différents contextes peuvent autoriser l'utilisation d'un système neuro-flou : les entrées ou les sorties pourraient être floues, le réseau neuronal pourrait être conçu comme un sous-système adaptatif d'un système flou, à défaut de valeurs précises et nettes, le réseau pourrait recourir à des nombres flous, etc. En général, comportent les avantages combinés de la logique floue et des réseaux de neurones : - Gèrent tous les types d'informations (linguistiques, logiques, numériques, …). - Prennent en charge les informations imparfaites et incomplètes, vagues et imprécises. - Permettent la résolution de conflits par aggrégation et collaboration. - Peuvent faire intervenir des procédures d'auto-calibration, d'auto-ajustement et d'auto-organisation. - Ne nécessitent pas la connaissance préalable des règles liant les données. - Miment les processus humains de prise de décision. - L'utilisation des opérations sur les nombres flous rendent les calculs très rapides.	Ce type d'hybridation constitue le cas classique d'un système regroupant les techniques d'apprentissage associées à l'observation des données et les techniques à base de connaissances : 1- on s'appuie d'abord sur les connaissances acquises du problème ainsi que sur l'expertises de spécialistes; 2- les connaissances préalables sont ensuite ajustées et calibrées aux données observées; 3- des règles floues sont extraites et validées par le réseau de neurones. On retrouve généralement deux utilisations des systèmes neuro-flous : (i) utiliser l'architecture d'un réseau neuronal pour simuler le système flou. (ii) utiliser un réseau de neurones pour adapter certains paramètres du système flou.
Neuro-	Les réseaux de neurones et les	Puisque les parties d'un réseau de

génétiques	algorithmes génétiques se combinent en dépassant les limites inhérentes à chacune de ces approches. Les algorithmes génétiques vont être utilisés surtout pour cloner les structures d'un réseau ou pour optimiser une rétropropagation dans un perceptron multicouche.	neurones peuvent aisément être regroupées, les algorithmes génétiques servent essentiellement à optimiser certains paramètres d'un réseau supervisé.
Flous-génétiques	Les algorithmes génétiques s'associent très bien à la logique floue, ce sont des techniques fortement robustes et complémentaires.	Les algorithmes génétiques permettent d'optimiser la base de règles floues.
Neuro-flous-génétiques	Ce sont des systèmes très complexes, mais lorsqu'ils sont bien adaptés aux contextes, s'avèrent très précis et viennent à bout des plus grandes difficultés.	Ce type de système hybride s'utilise dans des situations de grande complexité et soumises à de perpétuelles fluctuations, dans lesquelles la précision des sorties est particulièrement importante (évolution des marchés boursiers, modèles météorologiques, etc.).

Tableau 3 : tableau de l'hybridation de RNA, LF, AG. [26]

V. Exemple d'hybridation des Essaims de Particules avec le Recuit Simulé :

1. Définition de travail

Cette hybridation est proposée pour résoudre une problème dans le domaine de réseau électrique consiste à contrôler et à minimiser les pertes actives dans les réseaux électriques en utilisant une hybridation entre deux méthodes d'optimisations dont la première est la méthode d'optimisation par essaims de particules (OEP) et la deuxième est la méthode de recuit simulé .

2. des méthodes hybrides utilisées :

L'algorithme proposé effectue une recherche locale en utilisant la méthode du Recuit Simulé sur les 5% meilleurs membres de la population d'Essaims de Particules. Par contre, les autres opérateurs d'Essaims de Particules sont appliqués sur les 95% restant de la population. Pour chaque itération, les résultats obtenus à l'issue de l'application de cet algorithme hybride représentent la solution courante qui sera injectée au début de

l'itération suivante et le processus se reproduit jusqu'à l'obtention des meilleurs résultats, c'est à-dire la convergence

Organigramme des méthodes hybride utilisé :

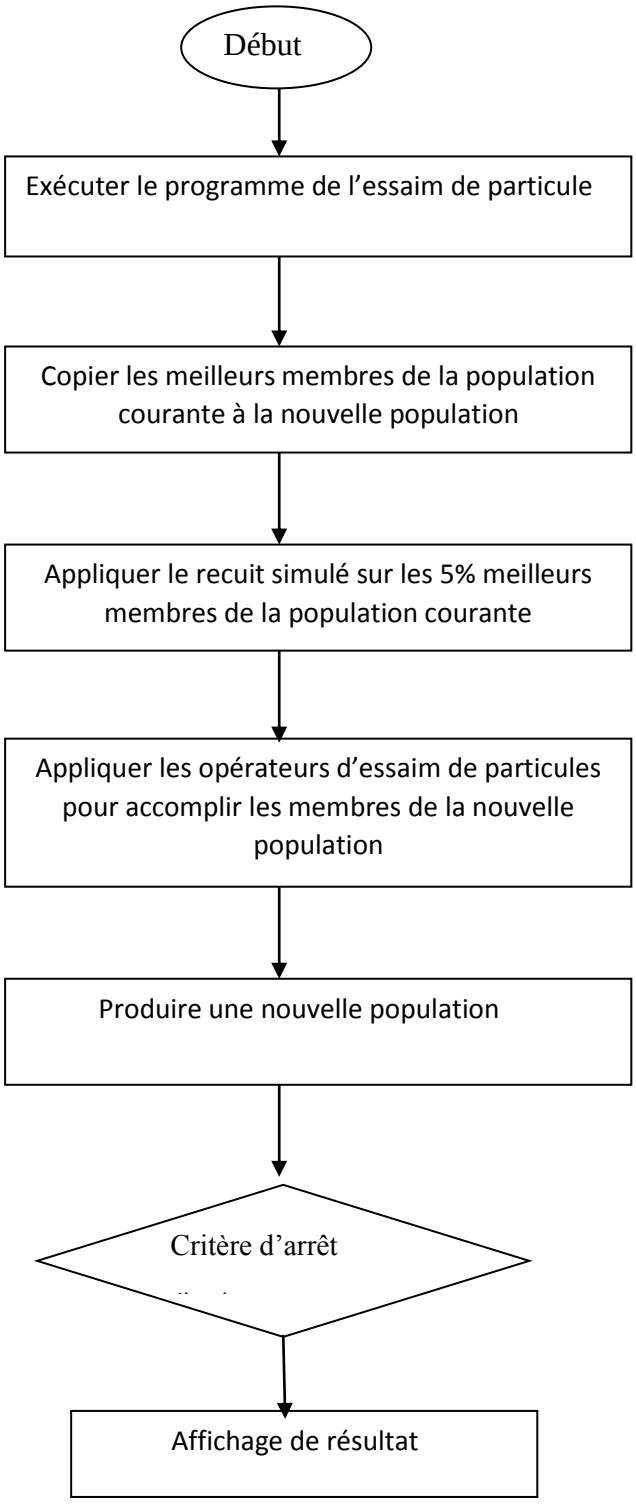

Figure 17 : Organigramme des méthodes hybride utilisé

3. Autres exemples d'hybridation de techniques :

Auteurs	Référence	Hybridation proposée	Principe de l'hybridation proposée	Domaine d'application
Li et al.	**[LI 08]**	**PSO -AG**	Après le classement des individus dans la génération actuelle, les meilleurs individus sont sélectionnés comme les élites en les reproduisant directement à la prochaine génération, les individus suivants sont évolués avec PSO, leurs meilleures positions sont mises à jour et les individus inférieurs sont évoluées avec une amélioration de l'algorithme génétique	Antennaarray pattern synthesis
Niknam et Amiri	**[NIK 10]**	**Fuzzyadaptive PSO-ACO et Kmeans**	La Sélection est pour chaque particule est en fonction de la méthode de sélection du meilleur chemin d'ACO. L'algorithme ACO-PSO est utilisé comme état initial de l'algorithme de K-means.	Problème de clustering partitional non linéaire (Nonlinear partitional clustering problem)
G.Kanagaj S.G.Ponambalam, N. Jawahar	**[ISSN0360-8352]**	**CS-GA**	-initialisation nombre de nids aléatoirement tel que chaque nids ayant un œuf correspondant une solution. Généré nouvelle population via les opérateurs d'algorithme génétique (sélection, croisement, mutation). Généré nouvelle solution via Lévy flight après l'évaluation fitness. Choisir une solution aléatoire et calculer sa fitness. Faire une comparaison entre les fitness et obtenu la meilleur solution.	A hybrid cuckoo search and genetical algorithm for reliability–redundancy Allocation problèmes

Tableau 4 : tableau de l'hybridation de PSO-ACO,PSO-AG,CS-AG . [27]

VI. la sélection d'attributs et les méthodes hybrides

1. généralité :

De nos jours, les chercheurs essaient de proposer des améliorations importantes qui permettent l'apparition d'une nouvelle génération de méthodes puissantes et générales. Ces dernières années, il est devenu évident que le fait de se concentrer sur une seule méthode pour résoudre un problème complexe est plutôt restrictif, une intégration de plusieurs approches peut améliorer l'efficacité et la flexibilité, surtout lorsqu'il s'agit de problèmes à grande dimension du monde réel. La tendance actuelle est l'émergence de méthodes hybrides pour la sélection d'attribut, qui essayent de tirer parti des avantages spécifiques de différentes approches en les combinant.[27]

2. Les méthodes hybrides utilisées pour la sélection d'attribut :

Les points suivant présentent quelques méthodes bio_inspéres hybrides pour la sélection d'attributs :

- **ACO_RNA:**

Une approche hybride basée sur l'ACO et les réseaux de neurones artificiels (RNA) pour trouver le sous-ensemble optimal de caractéristiques présentée dans [SIV 07]. Le modèle hybride proposée est évalué en utilisant des bases de données pour le diagnostic médical.[27]

- **PSO_AG :**

Dans [YAN 08], un algorithme binaire amélioré d'optimisation par les essaims particulaires a été intégré dans un algorithme génétique afin de servir comme optimiseur local pour chaque génération dans le problème de sélection des caractéristiques.[27]

- **PSO_AG:**

Basiri et Nemati [BAS 09] ont proposé un nouvel algorithme hybride ACO-AG pour la sélection d'attributs dans la catégorisation des textes. La performance de classification et la taille du sous-ensemble de caractéristiques sélectionnées sont adoptées comme des informations heuristiques.[27]

- **AG et ACO :**

Sheikhan et Mohammadi [SHE 12] ont développé un modèle hybride pour la prévision de charge à court terme. AG et ACO sont combinés dans ce modèle pour explorer l'espace de tous les sous-ensembles d'un ensemble de caractéristiques données, et le perceptron multicouche est utilisé pour la prédiction de la charge horaire. Sheikhan et Mohammadi [SHE 13] utilisent une méthode de sélection de caractéristiques AG-ACO hybride pour

obtenir le sous-ensemble de caractéristiques le plus petit et le plus efficace avec la base de données IEEE load dataset.[27]

- **BPNN-PSO**

Jin et al. [JIN 12] proposent sort-based BPNN-PSO (BPNN : Back Propagation Neural Network) pour sélectionner des attributs essentiels afin d'améliorer les performances de généralisation et réduire le coût de calcul de BPNN. Dans la méthode de sélection d'attributs proposée, la corrélation des entrées et des sorties est appliquée pour calculer l'importance des caractéristiques.[27]

- **MCS-SVM**

cette hybridation est une nouvelle technique pour la diagnostique de la malade de diabète qui est basé sur le « Feature Weighted Support Vector Machines et Modified Cuckoo Search » la sélection d'attribut est utilisé pour obtenu la classification des donnés, puis utiliser Modified cuckoo search pour trouver la valeur optimal pour les paramètres de SVM .[27]

VII. PSO-ICS-SVM pour la sélection d'attributs (Approche proposé) :

1. L'optimisation par essaim de particules (PSO) :

1.1. Définition :

L'optimisation par essaim de particules (le PSO en anglais : Particule Swarm Optimisation) est une métaheuristique à base de population de solution. Elle a été proposée en 1995 par Kennedy et Eberhart [Kennedy et Eberhart, 1995]. L'algorithme PSO est inspiré du comportement social d'animaux évoluant en essaim, tels que les poissons qui se déplacent en bancs ou les oiseaux migrateurs. En effet, on peut observer chez ces animaux des dynamiques de déplacement relativement complexes, alors qu'individuellement chaque individu a une intelligence limitée et une connaissance seulement locale de sa situation dans l'essaim.

La méthode d'optimisation par essaim particulaire met en jeu un ensemble d'agents pour la résolution d'un problème donné. Cet ensemble est appelé essaim. L'essaim est composé d'un ensemble de membres, ces derniers sont appelés particules. Les particules de l'essaim représentent des solutions potentielles au problème traité. L'essaim de particules survole l'espace de recherche, en quête de l'optimum global. Le déplacement de chaque particule est influencé par les trois composantes suivantes [Cooren, 2008].

a- Une composante physique : la particule tend à suivre sa direction de déplacement courante ;

b- Une composante cognitive : la particule tend à se diriger vers le meilleur site par lequel elle est déjà passée ;

c- Une composante sociale : la particule tend à se diriger vers le meilleur site déjà atteint par ses voisines.[28]

1.2. Implémentation :

Chaque particule i de l'essaim est définie par sa position $X_{id} =, X_{i2}, \dots, X_{id}, \dots, X_{iD})$ et sa vitesse de déplacement $V_{id} = (V_{i1}, V_{i2}, \dots, V_{id}, \dots, V_{iD})$ dans un espace de recherche de dimension D. Cette particule garde en mémoire la meilleure position par laquelle elle est déjà passée et la meilleure position atteinte par toutes les particules de l'essaim, noté respectivement :

$$P_{bestid=}(P_{besti1}, P_{besti2}, \dots, P_{bestid}, \dots, P_{bestiD})$$

$$et\ g_{bestid=}(g_{besti1}, g_{besti2}, \dots, g_{bestid}, \dots, g_{bestiD})$$

Le processus de recherche est basé sur deux règles :

- Chaque particule est dotée d'une mémoire qui lui permet de mémoriser la meilleure position par laquelle elle est déjà passée et elle a tendance à retourner vers cette position.
- Chaque particule est informée de la meilleure position connue au sein de son voisinage et elle a toujours tendance de se déplacer vers cette position.[28]

Début

1 : Initialiser les paramètres et la taille S de l'essaim ;

2 : Initialiser les vitesses et les positions aléatoires des particules dans chaque Dimension de l'espace de recherche ;

3 : Pour chaque particule, $P_{bestid} = X_{id}$;

4 : Calculer f (X_{id}) de chaque particule ;

5 : Calculer g_{bestid}; // la meilleure pbestid

6 : **Tant que** (la condition d'arrêt n'est pas vérifiée) **faire**

7 : **Pour** (i allant de 1 à S) **faire**

8 : Calculer la nouvelle vitesse ;

9 : Trouver la nouvelle position ;

10 : Calculer f (X_{id}) de chaque particule ;

11 : **Si** (f (X_{id}) est meilleur que f (P_{bestid})) **alors**

12 : $P_{bestid} = X_{id}$;

13 : **Si** (f (X_{id}) est meilleur que f (g_{bestid})) **alors**

14 : $g_{bestid} = P_{bestid}$;

15 : Fin pour

Algorithme 2 : Algorithme de L'optimisation par essaim de particules. [28]

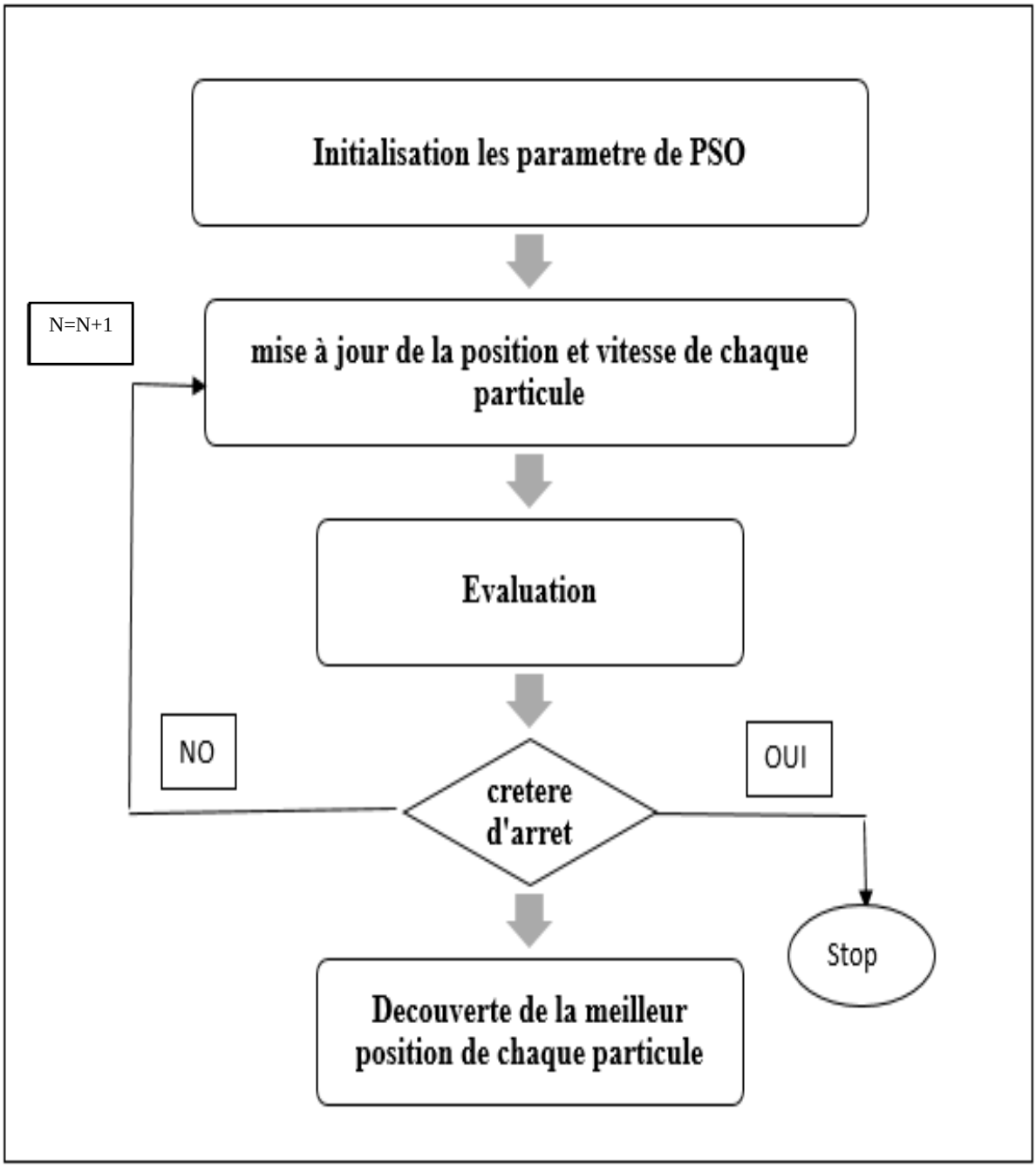

Figure 18 : diagramme de l'algorithme PSO

2. La recherche Coucou (CS)

2.1. Le principe et les étapes de la recherche coucou

En s'inspirant du comportement des coucous dans leur reproduction, Yang et Deb se sont basés sur trois principes pour proposer leur nouvelle métaheuristique [Yang et Deb, 2010].

-Chaque coucou pond un seul œuf à la fois. Il le dépose dans un nid qu'il choisit aléatoirement.

- Les meilleurs nids qui incluent des œufs (solutions) de bonnes qualités vont être les élus qui construisent les membres de la nouvelle génération.

-Le nombre des nids hôtes valides est fixé. L'oiseau hôte peut détecter le coucou étranger avec une probabilité $P_a \epsilon [0,1]$.Dans ce cas-là, l'oiseau hôte tranche entre écarter le coucou de son nid en lui éjectant hors nid ou abandonner son nid pour aller construire un autre dans une nouvelle position.

La probabilité P_a représente la fraction de N nids qui vont être remplacés par de nouveaux nids (avec de nouvelles solutions aléatoires dans de nouvelles positions dans l'espace de recherche). La qualité d'un nid ou d'une solution est mesurée en fonction de la fonction fitness qui se varie d'un problème à un autre.

Afin de générer une nouvelle solution $X_{(t+1)}$ pour un coucou i, Yang et Deb ont intégré le vol de Lévy de la manière suivante :

$$\mathrm{x_i}\ (\mathrm{t}+1) = x_i(t) + \alpha \oplus Lévy(\lambda)$$

Où $\alpha > 0$ est la taille du pas, elle est liée au problème traité.

La nouvelle solution sera donc générée en fonction de deux facteurs indispensables :

-La position actuelle du coucou.

-La nouvelle direction mesurée par le vol de Lévy.[28]

2.2. Le vol de Lévy

De nombreux phénomènes naturels ou sociaux, peuvent être décrits en termes de marche aléatoire. En général, le processus de recherche de nourriture chez les animaux est effectivement aléatoire. En fait, leur déplacement est basé sur leur position actuelle ainsi qu'une probabilité du déplacement vers une autre position. Des études expérimentales sur le comportement de certains animaux et insectes ont montré que leur comportement

peut être modélisé par un schéma mathématique nommé vol de Lévy (en anglais Lévy flight) [Brown et al 2007, Reynods and Frye 2007, Pavlyukevich 2007]. Le vol de Lévy ou Lévy flight a été proposé par le mathématicien français Paul Pierre Lévy, un des fondateurs de la théorie moderne de probabilités. Depuis sa création, le vol de Lévy a donné des interprétations théoriques à plusieurs phénomènes physiques, chimiques, biologiques et naturels. En fait, le vol de Lévy permet de modéliser des marches aléatoires composées d'un grand nombre de cas où les transitions sont basées sur des probabilités. En terminologie mathématique, le vol de Lévy est une marche aléatoire (en anglais, Random walk: une formalisation mathématique d'une trajectoire composée d'un ensemble de pas aléatoires) dans laquelle la distance entre les pas à une distribution probabilitaire (en anglais, probability distribution: une fonction qui représente la probabilité d'un nombre aléatoire de prendre une valeur donnée) à queue-lourde (en anglais, heavy-tail: dont les queues ne sont pas bornées de façon exponentielle) [Shlesinger et al 1995, Ben-Avraham et Havlin 2002].Plusieurs études récentes montrent qu'une panoplie de phénomènes dans différents domaines peuvent être modélisés par le vol de Lévy: Le déplacement des mouche s [Reynolds et Frye, 2007], la diffusion de la lumière [Barthelemy et al ,2008], Les mouvements des organismes biologiques [Viswanathana et al, 2002], la recherche aléatoire des objets [Viswanathana et al, 2000] et le domaine de l'optimisation et la recherche optimale où les études expérimentales ont montré des résultats encourageants [Shlesinger, 2006] [Pavlyukevich, 2007]. Restant dans le domaine de l'optimisation et des métaheuristiques,Yang et Deb [Yang et Deb, 2009] ont intégré de leur part le vol de Lévy dans leur récente métaheuristique: La recherche coucou (CS) pour générer de nouvelles solutions. [28]

2.3. Cuckoo search amélioré :

- **définition :**

La robustesse de CS est basée sur la manière d'explorer et d'exploiter l'espace des solutions par un coucou. Ce coucou peut avoir une certaine intelligence" pour trouver des solutions bien meilleures.

On considéré dans leur amélioration, le coucou comme le premier niveau de contrôle d'intersection et de diversification, et puisque ce coucou est un individu d'une population,

alors, cette population est qualifiée pour être le deuxième niveau de contrôle. L'idée de l'amélioration est de restructurer la population en intelligeant une nouvelle catégorie de coucous relativement plus intelligents avec plus d'efficacité dans leurs recherches, par rapport aux autres coucous.[29]

Les études ont montré que le coucou est capable d'engager une surveillance autour de potentiels nids hôtes [Payne et Sorenson (2005)].Ce comportement peut servir comme une inspiration de concevoir une nouvelle catégorie des coucous qui a une capacité de changer les nids hôtes durant l'incubation. Le but de ce comportement est d'éviter l'abandon des œufs des coucous. Ces coucous adoptent des mécanismes avant et après la couvaison. Ils observent le nid hôte choisi pour être sûr que le choix de ce nid est la bonne décision ou non (dans ce cas, ils commencent à chercher un nouveau choix bien meilleur que l'actuel).On parle, donc, d'une faculté de chercher localement une solution bien meilleure autour de la solution courante.[29]

Inspiré de ce comportement observe, le mécanisme adopté par cette nouvelle fraction de coucous, peut être divisé en deux étapes principales :

- Un coucou, initialement se déplace par les vols de Levy vers une nouvelle solution (qui représente une nouvelle région).
- A partir de la solution courante, un coucou dans la même région cherche une nouvelle meilleure solution (dans ce stade il est possible de procéder à une recherche locale).

Selon ces deux étapes, la population de l'algorithme CS amélioré peut être structurée suivant trois principales catégories de coucous :

-Un coucou, cherchant (à partir de la meilleure position) des régions pouvant contenir de nouvelles solutions qui sont bien meilleures que la solution d'un individu sélectionné aléatoirement dans la population ;

-Une fraction **Pa** des coucous cherchant de nouvelles solutions loin de la meilleure solution ;

-Une fraction **pc** des coucous cherchant des solutions à partir de la position courante et essayant de les améliorer. Ils se déplacent d'une région vers une autre via les vols de Levy pour localiser la meilleure solution dans chaque région sans être pièges par l'optimum local.

Nous notons que la population des coucous, durant son processus de recherche, est guidée par : **la meilleure solution, les solutions trouvées localement, et les solutions trouvées loin de la meilleure solution.** Ceci, améliore la recherche intensive autour des différentes meilleures solutions, et en même temps, une randomisation est proprement

réalisée an d'explorer de nouvelles régions à l'aide des vols de Levy. Ainsi, une extension de la version standard de CS, est l'ajout d'une méthode qui manipule la fraction **pc** des coucous intelligents. Ceci permet à CS de fonctionner plus efficacement avec peu d'itérations, et montrer plus de résistance face aux potentiels pièges ou stagnations dans les optimums locaux.

Le nouveau processus ajouté à CS standard peut être illustré par une procédure considérant que la valeur de la fraction **pc** est fixée à 0.6, telle que cette fraction est un ensemble de bonnes solutions de la population sauf la meilleure solution. Partant de chaque solution, un coucou effectue une recherche à pas aléatoires via les vols de Levy autour de la solution courante, et puis il essaye de trouver la meilleure solution dans cette région en utilisant la recherche locale. [29]

Le but de cette amélioration est de renforcer la recherche intensive autour des meilleures solutions de la population, tout en considérant la randomisation qui doit être guidée proprement par les vols de Levy pour l'exploration de nouvelles régions. En effet, une extension de CS standard est l'ajout d'une méthode qui gère la fraction **pc** des coucous intelligents.

On peut dire que la nouvelle catégorie des coucous permet à CS d'être plus performant et plus efficace avec moins d'itérations. Elle offre un type de résistance face aux blocages éventuels dans les optimums locaux.

- **Implémentation** :

La principale différence entre l'ICS et CS est dans les valeurs de Pa et α. Pouramélioreria performance de l'algorithme CS et éliminer les inconvénients relié avec les valeurs fixé de Pa et α, l'algorithme ICS doivent utiliser des variables Pa et α.Dans les premières générations, les valeurs de Pa et α doit être assez grand pour appliquer l'algorithme pour augmenter la diversité des vecteurs de la solution. Toutefois, ces valeurs doivent être réduites dans les générations finales pour aboutir à un meilleur **réglage-fine-tuning-** des vecteurs de la solution. Les valeurs de Pa et α sont modifiées dynamiquement, le nombre de génération est exprimé dans les équations .1.3, où **NI** et **gn** sont le nombre d'itérations total et l'itération courante, respectivement. [29]

$$P_a\,(gn) = P_{a\,max} - \frac{gn}{NI}\,(P_{a\,max} - P_{a\,min}) \quad .1$$

$$\alpha(gn) = \alpha_{max} e^{c.gn} \quad .2$$

$$c = \frac{1}{NI}\ln(\frac{\alpha_{min}}{\alpha_{max}}) \quad .3$$

2.4. L'algorithme de la recherche coucou amélioré(ICS) :

1 : Fonction objectif f(x), x=$(x_1, \ldots\ldots, x_d)^t$

2 : Générer la population initiale de n nids hotex_i (i = 1,…. n)
3 : tantque (t<MaxGénération) ou (critère d'arrêt) faire
4 : lancer la recherche avec une fraction (p_c) des coucous intelligents
5 : Obtenir un coucou aléatoire par le vols de lévy
6 : Evaluer sa qualité/fitness F_i
7 : choisir un nid parmi n (soit, j) aléatoirement
8 : si ($F_i > F_j$) alors
9 : remplacer j par la nouvelle solution ;
10 : finsi
11 : une fraction (p_a) des mauvais nids est abandonné et des nouveaux sont construits ;
12 : Garder les meilleurs solutions (ou nids avec des solutions de qualité) ;
13 : classer les solutions et trouver la meilleure actuelle
14 : fin tantque
15 : Post_processus des résultats et visualisation

Algorithme 3 :L'algorithme de la recherche coucou amélioré(ICS)

L'idée de l'amélioration introduite dans l'algorithme CS est la recherche des solutions dans des régions spécifiées par les vols de Levy indépendamment de la meilleure solution dans la population. On constate aussi que cette amélioration se présente comme une variante de la recherche locale autour d'une fraction pc des solutions. L'unique inconvénient de la recherche locale est le blocage dans les optimums locaux. Cet

inconvénient est bien traite dans le cas CS améliore qui exige un déplacement par région et non pas par solution, ce qui minimise remarquablement ces blocages.[29]

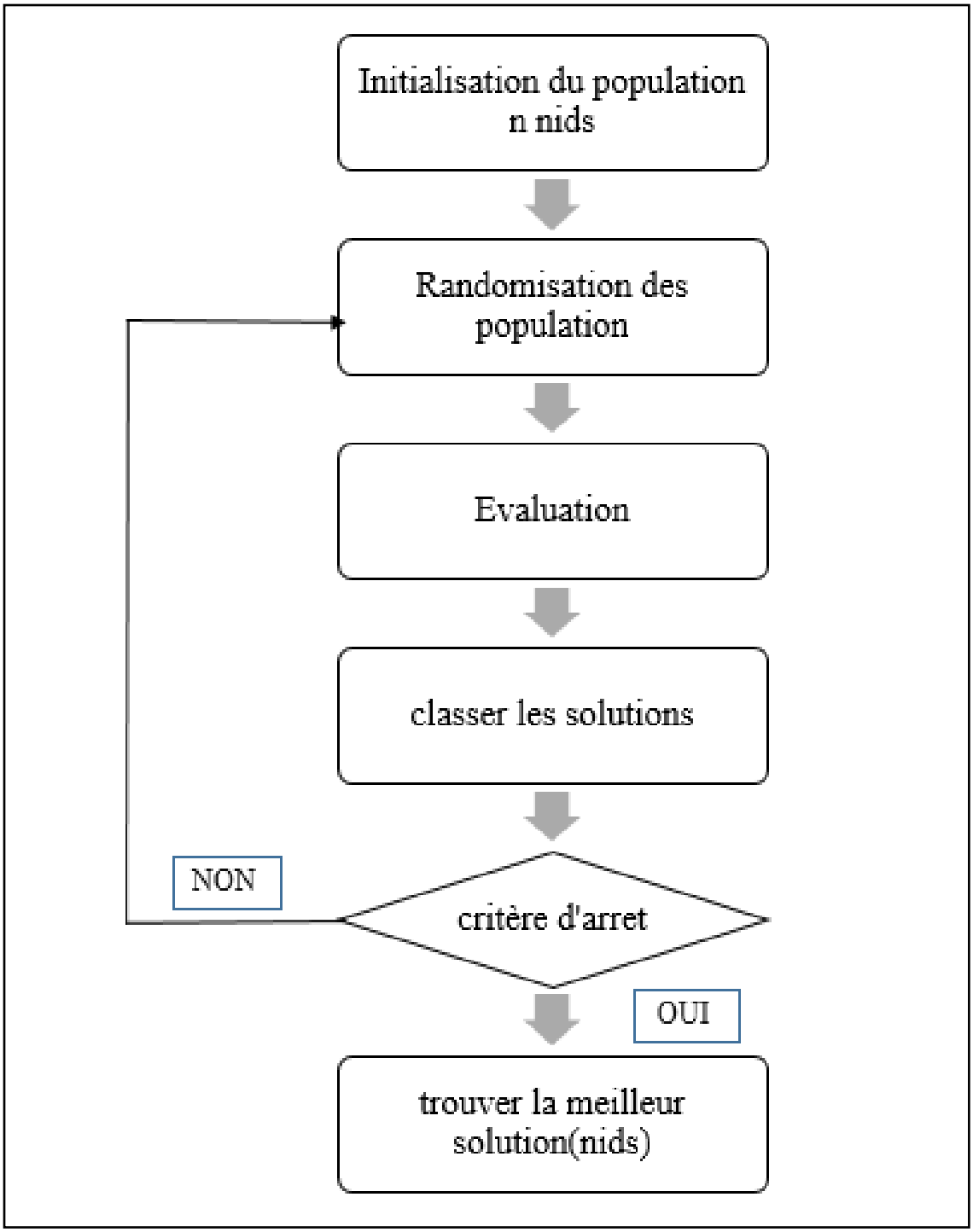

Figure 19 : diagramme de l'algorithme ICS

3. SVM : Machines à vecteurs de Support ou Séparateurs à Vaste Marge :

3.1. Général :

Parmi les méthodes à noyaux, inspirées de la théorie statistique de l'apprentissage de Vladimir Vapnik, les SVM constituent la forme la plus connue. SVM est une méthode de classification binaire par apprentissage supervisé, elle fut introduite par Vapnik en 1995. Cette méthode est donc une alternative récente pour la classification. Cette méthode repose sur l'existence d'un classificateur linéaire dans un espace approprié. Puisque c'est un problème de classification à deux classes, cette méthode fait appel à un jeu de données d'apprentissage pour apprendre les paramètres du modèle. Elle est basée sur l'utilisation de fonction dites noyau (kernel) qui permettent une séparation optimale des données.

Dans la présentation des principes de fonctionnements, nous schématiserons les données par des « points » dans un plan. La notion d'apprentissage étant importante, nous allons commencer par effectuer un rappel. L'apprentissage par induction permet d'arriver à des conclusions par l'examen d'exemples particuliers. Il se divise en apprentissage supervisé et non supervisé. Le cas qui concerne les SVM est l'apprentissage supervisé. Les exemples particuliers sont représentés par un ensemble de couples d'entrée/sortie. Le but est d'apprendre une fonction qui correspond aux exemples vus et qui prédit les sorties pour les entrées qui n'ont pas encore été vues. Les entrées peuvent être des descriptions d'objets et les sorties la classe des objets donnés en entrée.[30]

3.2. SVM principe de fonctionnement général :

- **Notions de base : Hyperplan, marge et support vecteur**

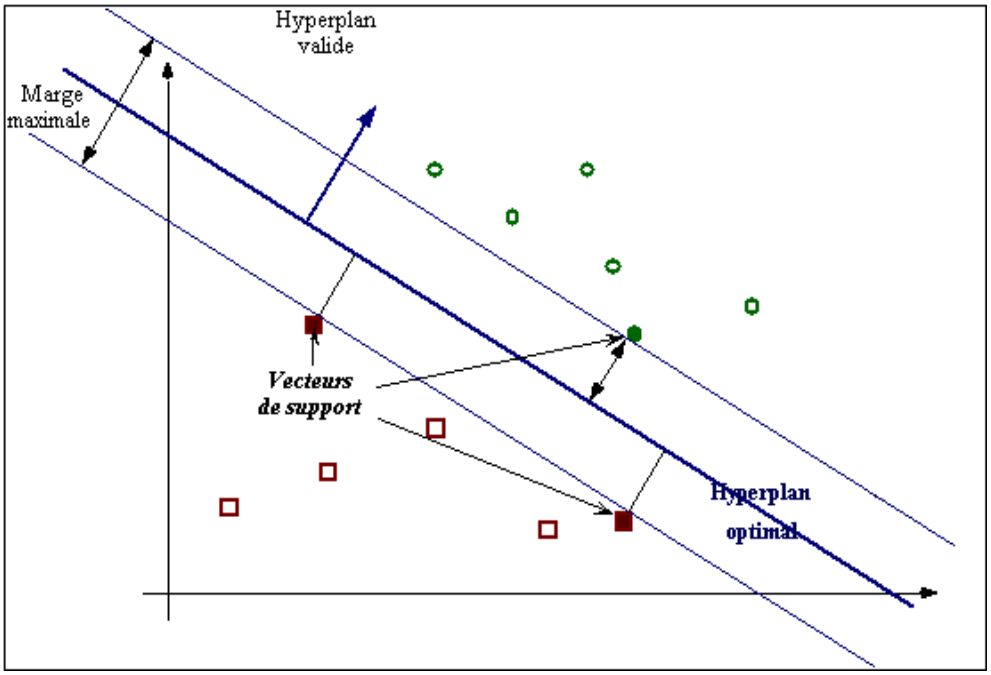

Figure 20 : Hyperplan, marge et support vecteur. [30]

- **vecteurs de support** : Les points les plus proches, qui seuls sont utilisés pour la détermination de l'hyperplan.
- **La marge et l'hyperplan :**

Il est évident qu'il existe une multitude d'hyperplan valide mais la propriété remarquable des SVM est que cet hyperplan doit être optimal. Nous allons donc en plus chercher parmi les hyperplans valides, celui qui passe « au milieu » des points des deux classes d'exemples. Intuitivement, cela revient à chercher l'hyperplan le « plus sûr ». En effet, supposons qu'un exemple n'ait pas été décrit parfaitement, une petite variation ne modifiera pas sa classification si sa distance à l'hyperplan est grande. Formellement, cela revient à chercher un hyperplan dont la distance minimale aux exemples d'apprentissage est maximale. On appelle cette distance « ***marge*** » entre l'hyperplan et les exemples. L'hyperplan séparateur optimal est celui qui maximise la marge. Comme on cherche à maximiser cette marge, on parlera de ***séparateurs à vaste marge.***[30]

- **Pourquoi maximiser la marge ?**

Intuitivement, le fait d'avoir une marge plus large procure plus de sécurité lorsque l'on classe un nouvel exemple. De plus, si l'on trouve le classificateur qui se comporte le mieux vis-à-vis des données d'apprentissage, il est clair qu'il sera aussi celui qui permettra au mieux de classer les nouveaux exemples. Dans le schéma qui suit, la partie droite nous montre qu'avec un hyperplan optimal, un nouvel exemple reste bien classé alors qu'il tombe dans la marge. On constate sur la partie gauche qu'avec une plus petite marge, l'exemple se voit mal classé.

En général, la classification d'un nouvel exemple inconnu est donnée par sa position par rapport à l'hyperplan optimal. Dans le schéma suivant, le nouvel élément sera classé dans la catégorie des « + ».[30]

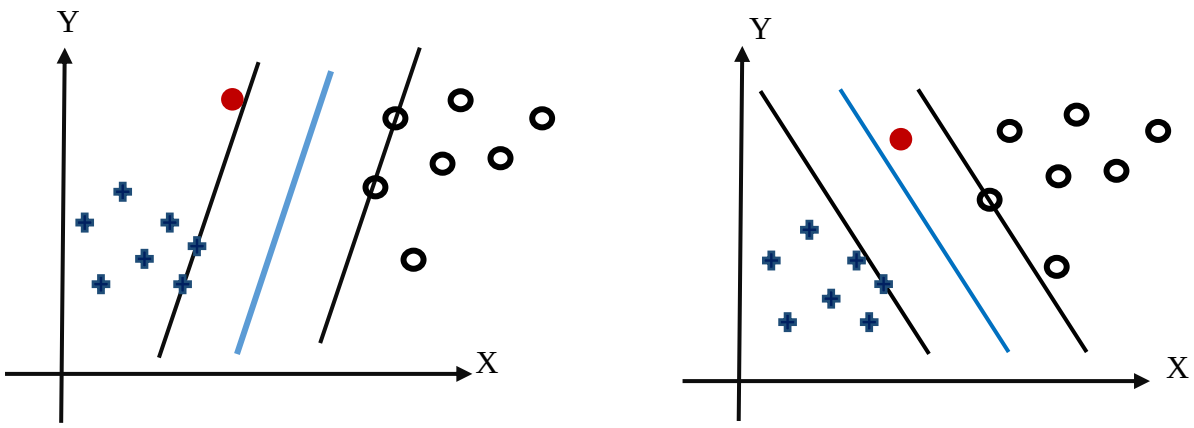

Figure 21 :maximisation de la marge

- **Linéarité et non-linéarité**

Parmi les modèles des SVM, on constate les cas linéairement séparable et les cas non linéairement séparable. Les premiers sont les plus simples de SVM car ils permettent de trouver facilement le classificateur linéaire. Dans la plupart des problèmes réels il n'y a pas de séparation linéaire possible entre les données, le classificateur de marge maximale ne peut pas être utilisé car il fonctionne seulement si les classes de données d'apprentissage sont linéairement séparables.[30]

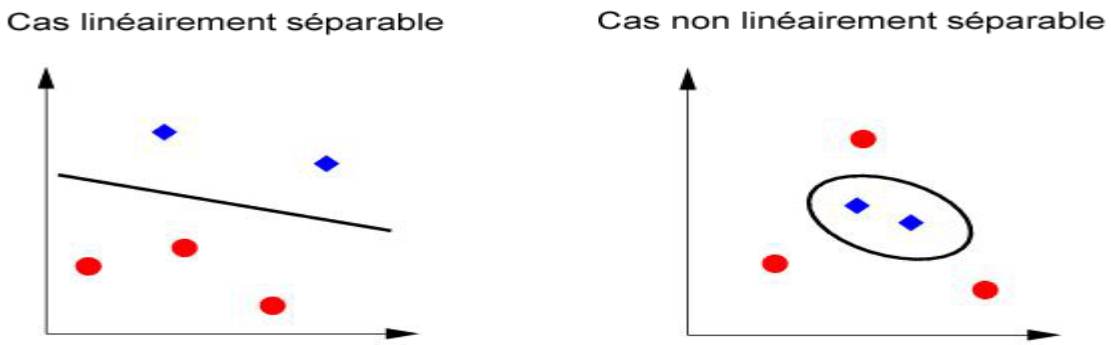

Figure 22 : les cas linéairement séparable et les cas non linéairement séparable

- **Cas non linéaire :**

Pour surmonter les inconvénients des cas non linéairement séparable, l'idée des SVM est de changer l'espace des données. La transformation non linéaire des données peut permettre une séparation linéaire des exemples dans un nouvel espace. On va donc avoir un changement de dimension. Cette nouvelle dimension est appelé « espace de re-description ». En effet, intuitivement, plus la dimension de l'espace de re-description n'est grande, plus la probabilité de pouvoir trouver un hyperplan séparateur entre les exemples est élevée. Ceci est illustré par le schéma suivant :

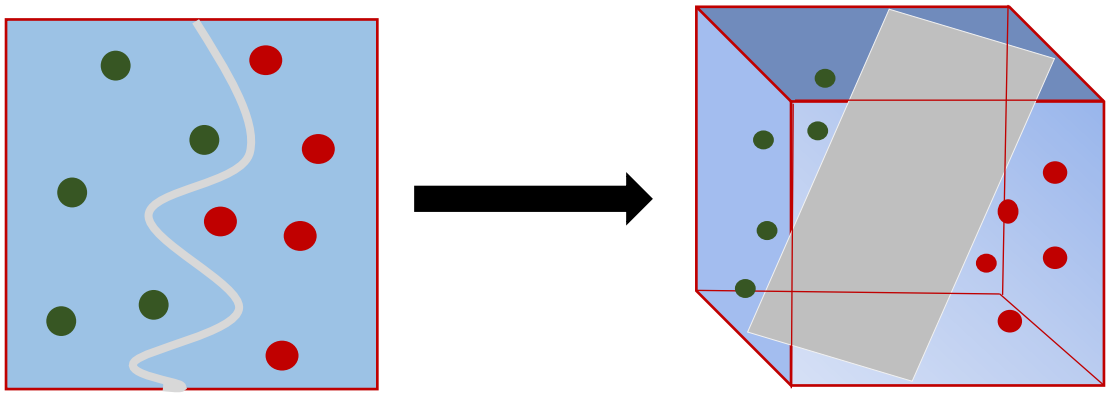

Figure 23 : Espace de re-description

On a donc une transformation d'un problème de séparation non linéaire dans l'espace de représentation en un problème de séparation linéaire dans un espace de re-description de plus grande dimension. Cette transformation non linéaire est réalisée *via* une fonction noyau. En pratique, quelques familles de fonctions noyau paramétrables sont connues et il revient à l'utilisateur de SVM d'effectuer des tests pour déterminer celle qui convient le mieux pour son application. On peut citer les exemples de noyaux suivants : polynomiale, gaussien, sigmoïde et Laplacien.[30]

$$\Phi : R^d \rightarrow \boldsymbol{\mathcal{F}}$$

$$\boldsymbol{x} \rightarrow \boldsymbol{\Phi}(\boldsymbol{x})$$

3.3. Fonction noyau (kernel) :

Dans le cas linéaire, on pouvait transformer les données dans un espace où la classification serait plus aisée. Dans ce cas, l'espace de redescription utilisé le plus souvent est ***R*** (ensemble des nombres réels). Il se trouve que pour des cas non linéaires, cet espace ne suffit pas pour classer les entrées. On passe donc dans un espace de grande dimension.

Avec card(***F***) > d.[30]

- **Exemples de noyaux :**
- Linéaire $k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x} \cdot \boldsymbol{x}'$
- Polynomial $k(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x} \cdot \boldsymbol{x}')^d$ ou $(c + \boldsymbol{x} \cdot \boldsymbol{x}')^d$
- Gaussien $k(\boldsymbol{x}, \boldsymbol{x}') = e^{-\|\boldsymbol{x}-\boldsymbol{x}'\|^2/\sigma}$
- Laplacien $k(\boldsymbol{x}, \boldsymbol{x}') = e^{-\|\boldsymbol{x}-\boldsymbol{x}'\|_1/\sigma}$

4. L'hybridation de ICS_PSO_SVM proposée :

Les étapes de l'algorithme ICS_PSO_SVM sont comme suit :

1-Initialiser la population de PSO ;

2-Initialise vitesse et position pour n particules ;

3-Evaluer chaque particule /fitness ;

4-Obtenir la solution global et locale de chaque particule ;

5-Utiliser les solutions d'algorithme de PSO comme population initial d'ICS

6- Evaluation chaque nid/fitness ;

7-Modifier la position de nid ;

8-Critère d'arrêt ;

9-Trouver la meilleur solution ;

10-validation par SVM

Diagramme de l'hybridation proposé ICS_PSO_SVM :

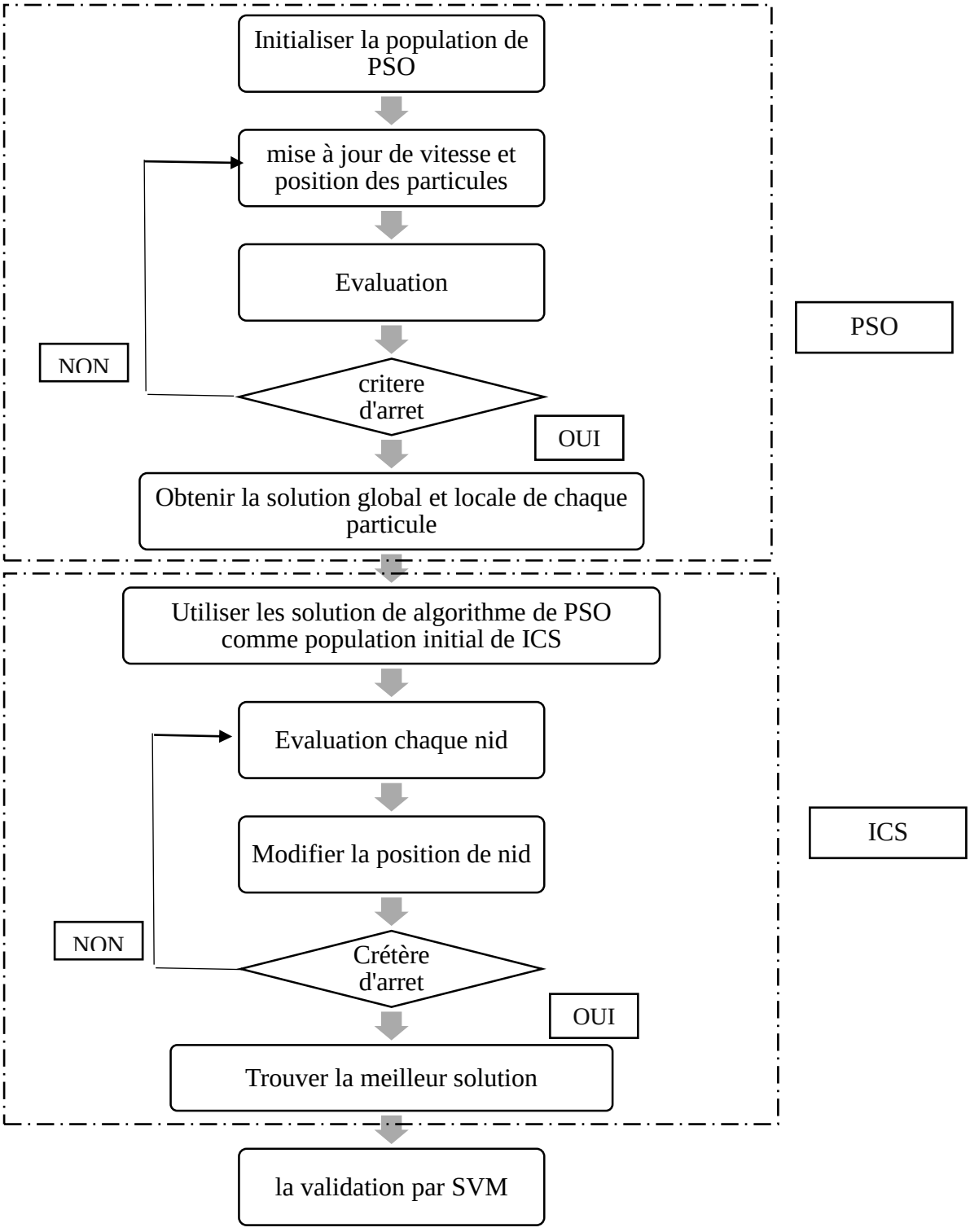

Figure 24 :Diagramme de l'hybridation proposé ICS_PSO_SVM

3.5 La méthode de sélection pour l'approche proposée :

Pour notre approche proposée, nous allons utiliser la méthode de sélection par l'étape de validation (Les méthodes enveloppantes) –wrrapper-, le diagramme suivant représente la méthode utilisé :

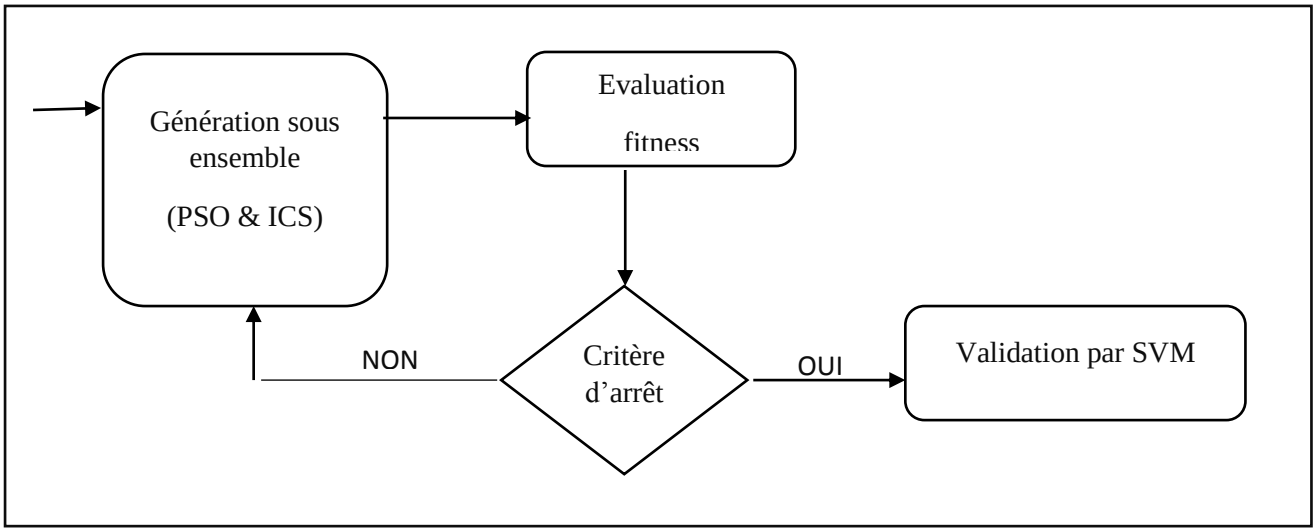

Figure 25 : Processus de la sélection des attributs avec l'étape de validation pour l'approche proposée.

VIII. Conclusion :

Ce chapitre est constitué de trois parties, dans la première partie nous avons expliqué brièvement les méthodes d'hybridation avec leur classification et présenté quelques exemples, la deuxième partie évoque les méthodes hybrides déjà utilisées pour la sélection d'attributs. Dans la troisième partie, nous avons présenté l'approche proposée (PSO-ICS-SVM) pour la sélection d'attributs.

Chapitre IV :

Implémentation et Résultats expérimentaux

I. Introduction :

Dans ce chapitre, nous présentons le Matériel ainsi le langage de programmation utilisés, Des données ont été utilisées pour valider l'approche proposée, Les tests utilisés, et les résultats obtenus sont présentés.

II. Outils de travail :

II.1. Le matériel :

Nous avons réalisé notre application sur un microordinateur de RAM 3 Go, Disque dur320Go, Processeu**r** pentium (R) Dual- corecpuT4500 2 ,30GHzet le système exploitation est Windows 8.1.

II.2. Le langage utilisé :

Pour implémenter notre application, nous avons choisi le langage Matlab version 7.12.0 (R2011a) .D'abord, c'est un outil réputé et utilisé dans plusieurs domaines des sciences de l'ingénieur (dont Traitement du Signal et des Images, Optimisation, Automatique), ensuite Matlab est un outil qui permet de faire des calculs numériques d'algèbre linéaire (en gros, les vecteurs, les matrices et tout ce qui s'y rapporte) de façon assez simple sur des quantités de données relativement conséquentes. Pour dire ça de façon simpliste (et réductrice) :

- C'est plus compliquer à utiliser qu'Excel mais c'est (bien) plus puissant.
- C'est moins performant qu'un truc que nous allons coder en C, mais c'est beaucoup plus simple à mettre en œuvre.

Aussi ce langage évolue nos recherches et nous fait gagner du temps grâce à ses fonctions prédéfinies.

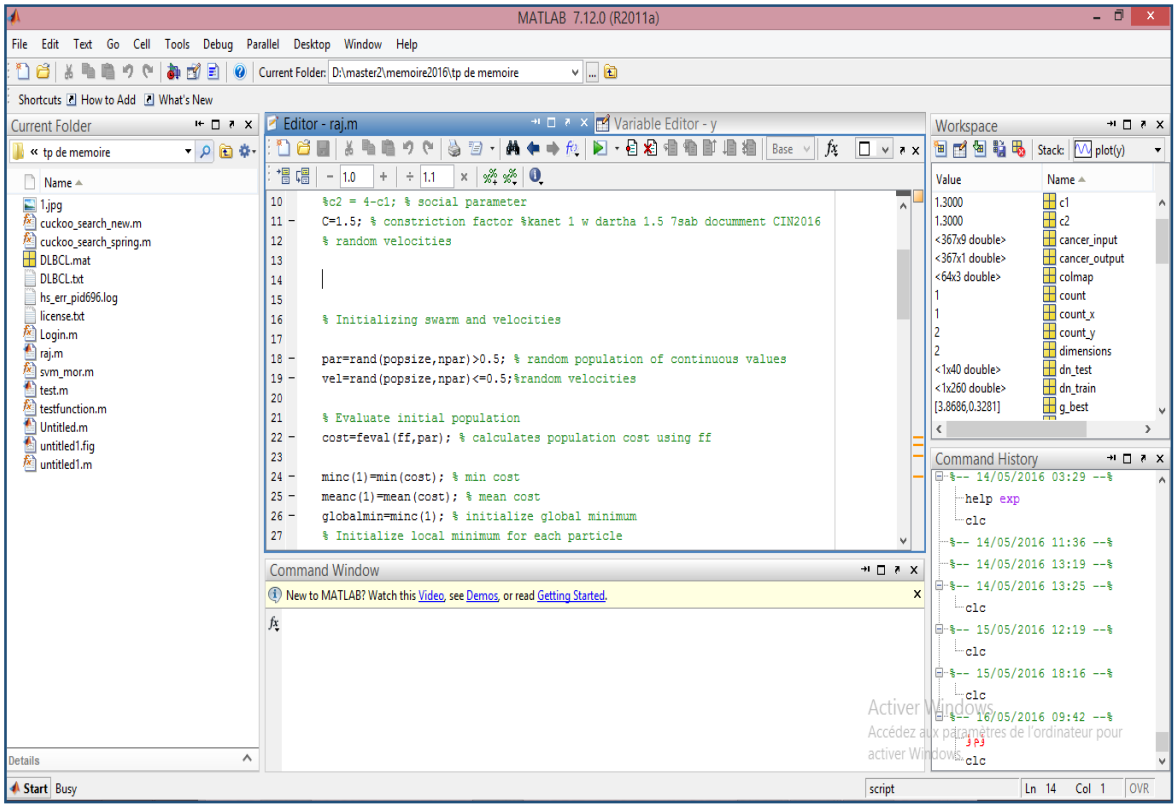

Figure 26 : Fenêtre principale de MATLAB2011

II.1.3 Les données utilisées :

Notre étude porte sur la sélection des attributs supervisés, depuis plusieurs ensembles de données de puces d'ADN ont des valeurs de classe qui sont utiles pour la prédiction.

Pour tester l'efficacité de l'approche proposée, un ensemble de données de puces à ADN sont utilisés [39]. Les ensembles de données ont été recueillies a partir de Kent Ridge bio-médical data repository[1] et Gene Expression model selection[2] , à partir Université Vanderbilt. Les principales caractéristiques des ensembles de données sont présentées dans le tableau 5.

[1] http://datam.i2r.a-star.edu.sg/datasets/krbd/
[2] http://www.gems-system.org/

Données	# Attributs	# Classes	# Echantillon
Prostate _ Tumor	10 509	2	102
Colon	2 000	2	62
Leukemia	7 129	2	72
Ovarian	15 154	2	253
DLBCL	5 469	2	77

Tableau 5: Table des bases de données utilisées.

III. Les paramètres de l'approche proposée :

Comme nous avons déjà signalé dans les chapitres précédents, l'approche proposée est une hybridation de trois algorithmes ; deux algorithmes d'optimisation qui sont : l'algorithme particule swarm optimisation PSO, et l'algorithme de recherche coucou amélioré (ou improved cuckoo search en anglais) ICS, avec le classifieur SVM, les tableaux suivants représentent les paramètres utilisées pour chacun des algorithmes.

III.1 Les paramètre de l'algorithme SVM :

Paramètre	Valeur
Type de model	SVM (support vector machine)
La fonction Karnel SVM	RBF (Radial basis)
La méthode de validation	K-fold cross validation
Le nombre de fold cross validation (K)	15

Tableau 6: Table des paramètres de SVM

III.1 Les paramètre de l'algorithme PSO :

Paramètre	Valeur
La taille de l'essaim	15
La dimension de problème	2
Nombre maximal des itérations	99
Le paramètre cognitif et social	c1 = 2 ; c2 = 2 ;

Facteur de construction	C=1.5 ;
La masse d'inertie	W=0.7

Tableau 7 : Table des paramètres de PSO

III.1 Les paramètre de l'algorithme ICS :

Paramètre	**Valeur**
Nombre des nids (ou les solutions diffèrent)	N=15
Nombre total d'itération	N_IterTotal=100 ;
Taux découvert les nids /solution	pamax = 1 ; pamin = 0.005 ;
nombre de domaine	Nd=25
Stepsize Beta	betamin=0.05 ; betamax=0.5 ;

Tableau 8: Table des paramètres d'ICS

IV. Les interfaces de l'application :

Figure 27 : Interface d'authentification

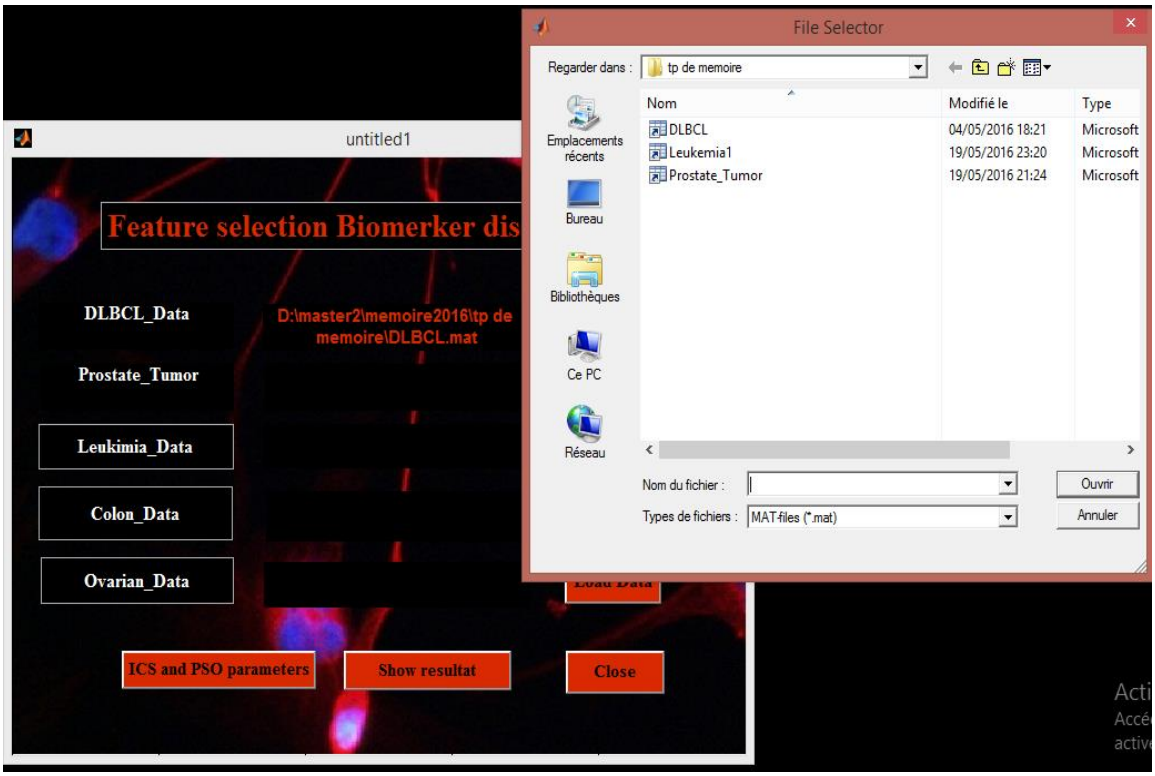

Figure 28 : Interface de charger le Data

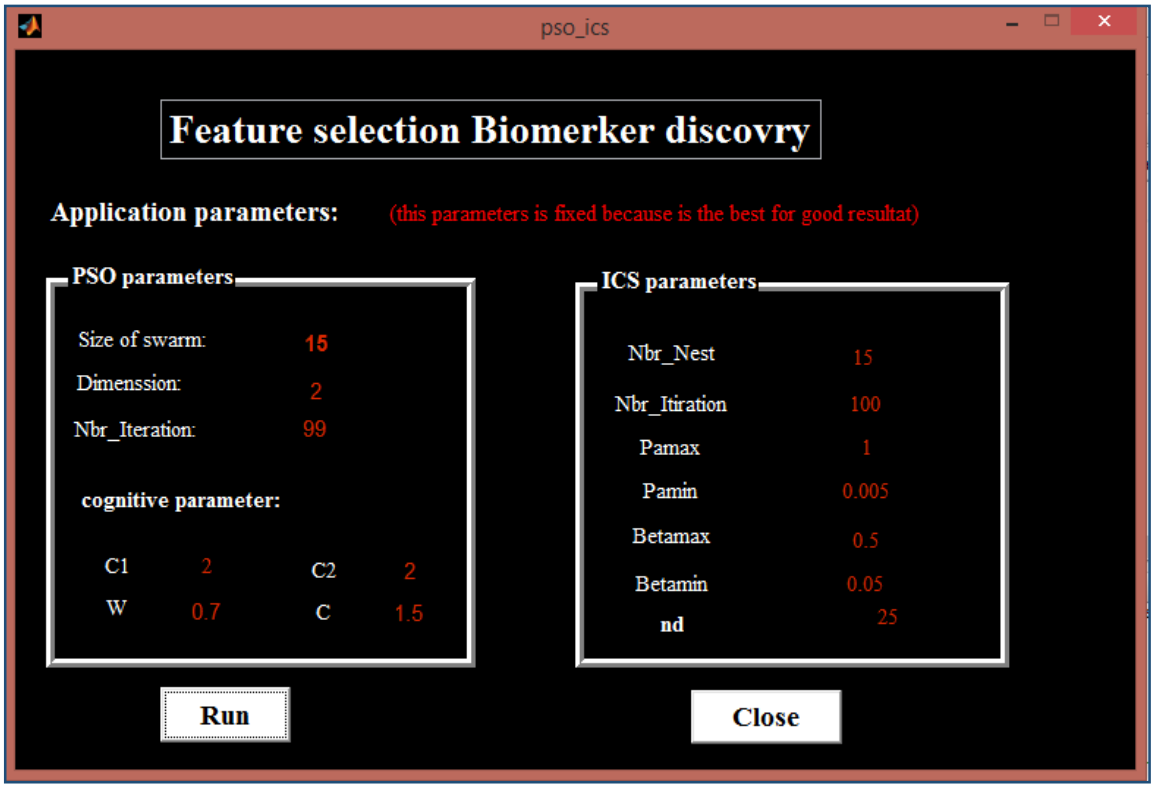

Figure 29 : Interface des paramètres de PSO et ICS

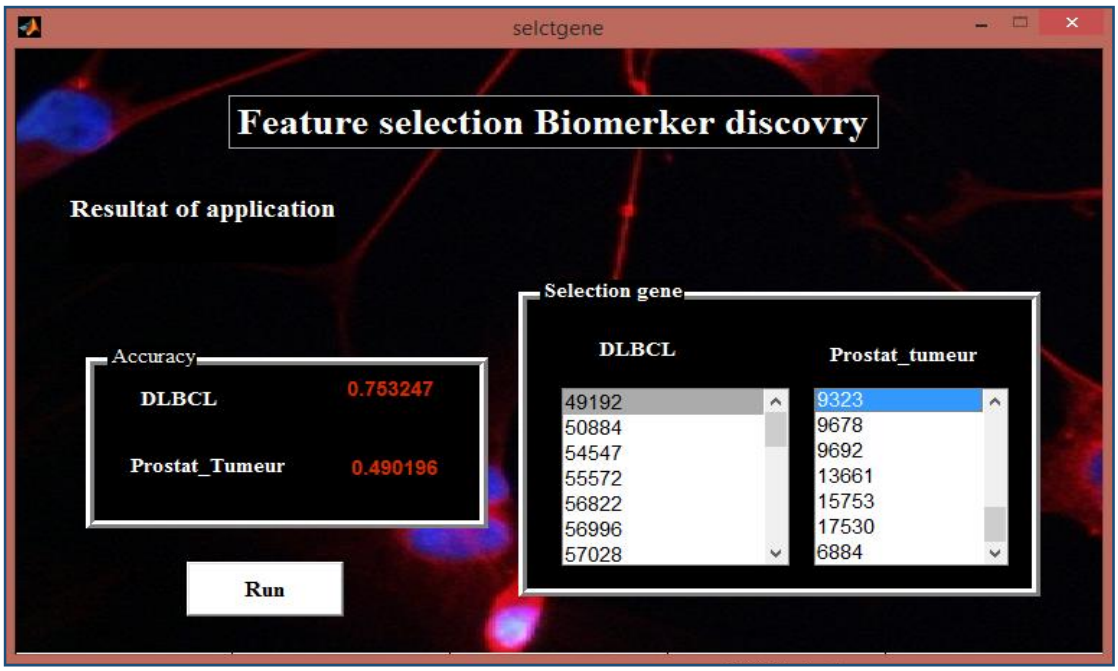

Figure 30 : Interface de résultat de sélection gènes

Les meilleurs 25 des gènes sélectionnés obtenus avec PSO_ICS_SVM :

Prostate_Tumeur = 0.5
7064 7081 7299 7777 7800 8081 8097 8152 8248 8369 8397 8473 8839 9044 9196 9220 9323 9678 9692 13661 15753 17530 6884 6933 7016

DLBCL=0.75
49192 50884 54547 55572 56822 56996 57028 58376 59037 60997 61729 62270 63096 63630 67734 71933 75234 84063 46429 46446 46663 47312 48067 48202

Tableau 9 : Table des meilleurs 25 gènes de DLBCL et Prostat_Tumeur data

V. Conclusion :

Dans ce chapitre, nous avons présenté les résultats de l'approche proposée qui consiste en l'hybridation de trois d'algorithmes utilisée pour la découverte de biomarqueurs.

Les résultats obtenus montrent l'avantage de l'hybridation des algorithmes d'optimisation avec les algorithmes d'apprentissage automatique.

Conclusion Générale et Perspectives

Conclusion générale et perspectives :

I. Conclusion générale :

L'évolution rapide de l'informatique fut associée à une montée éclatante de problèmes informatiques de tous genres. La technologie des ordinateurs s'est développée à un rythme effréné, rendant possible le traitement de problèmes de plus en plus compliqués. Mais les problèmes sont devenus tellement complexes qu'il n'est plus possible à l'être humain de compter uniquement sur la puissance de son ordinateur ni même sur celle de son super calculateur.

La sélection des caractéristiques ou attributs est l'une des étapes les plus importantes dans le traitement des données dans différents domaines applicatifs. Elle consiste à trouver un sous-ensemble résoudre des problèmes complexes divers, dans plusieurs domaines de recherche et d'ingénierie.

Le travail de ce mémoire a mis l'accent sur l'utilité de l'hybridation des méthodes (hybridation entre l'optimisation PSO, ICS et l'apprentissage automatique SVM) pour la découverte des biomarqueurs à partir des puces à ADN dans le domaine de la bioinformatique.

L'apport de l'hybridation est totalement bénéfique et résultats satisfaisants.

II. Perspectives

Au terme de ce travail, il parait tout à fait intéressant de suggérer des études complémentaires. En particulier de généraliser l'approche PSO-ICS-SVM pour d'autres problèmes, et utilisé d'autres hybridation pour des problèmes dans le domaine de la bioinformatique.

Aussi, une comparaison est envisagée entre les différentes méthodes qui existent pour bien extraire les avantages de chacune. Le résultat peut servir à une combinaison qui peut conduire à des taux de succès plus élevé.

Bibliographie

9. LABED SAID «Méthodes bio-inspirées hybrides pour la résolution de problèmes complexes », thèse Doctorat en Sciences en Informatique, Université Constantine 2,2013.

10. Documents Electroniques, Optimisation Combinatoire (Méthodes approchées), présentation par les membre de laboratoire d'informatique fondamentale de Lille (LIFL),2008/2009.

11. S. Ben Ismail « Introduction à l'optimisation combinatoire », Majeure Informatique – INF413 – C5 ,2 eme semestre 2012 .

12. MAHDI SAMIR « Optimisation Multiobjectif Par Un Nouveau Schéma De Coopération Méta/Exacte », Mémoire de Magister Spécialité : Intelligence Artificielle et Génie Logiciels, Université Mentouri de Constantine.

13. Alaoui Abdiya « Application des techniques des métaheuristiques pour l'optimisation de la tâche de la classification de la fouille de données », mémoire Pour l'obtention du diplôme de Magister en informatique, université des sciences et de la technologie D'ORAN Mohamed Boudiaf, 2011/2012.

14. Documents Electroniques, Optimisation combinatoire et meaheuristiques, IFT1575 Modèles de recherche opérationnelle (RO).

18. Gael Le Mahec « Gestion des bases de données biologiques sur grilles de calcul » ,2008.

21. Bendana Rokia « Sélection d'Attributs Basée sur un Algorithme Génétique Neuronal : Application à la Reconnaissance des Caractères Manuscrits », Option intelligence artificielle et génie logiciel, 2007.

23. Sabra el Ferchichi, « traitement de signal et image », génie informatique.

24. Bendana Rokiya, « Sélection d'Attributs Basée sur un Algorithme Génétique Neuronal : Application à la Reconnaissance des Caractères Manuscrits >>, Option intelligence artificielle et génie logiciel, 2007.

25. Hanaa Hachimi. « Hybridations d'algorithmes metaheuristiques en optimisation globale et leurs applications. », Ecole Mohammadia d'ingénieurs (Rabat, Maroc), 2013.

27. MENGHOUR Kamilia , « Approches Bio-inspirées pour la Sélection d'Attributs », BADJI MOKHTAR-ANNABA UNIVERSITY,2015

28. Amira Gherboudj Méthodes de résolution de problèmes difficiles académiques, thèse doctorat, Université de Constantine2, 2013.

29. Aziz OUAARAB, « Résolution de Problèmes d'Optimisation Combinatoire par des Métaheuristiques Inspirées de la Nature : Recherche du Coucou via les Vols de Lévy », UNIVERSITÉ MOHAMMED V FACULTÉ DES SCIENCES, Rabat, 2015.

30. Mohamadally Hasan et Fomani Boris, « SVM : Machines à Vecteurs de Support ou Séparateurs à Vastes Marges », BD Web, ISTY3, Versailles St Quentin, France, ,16 janvier 2006.

39. Martinez, E., et al. « Compact cancer biomarkers discovery using a swarm intelligence feature selection algorithm. Computational Biology and Chemistry », 34, 244–250 (2010)

WEBGRAPHIE

1. https://fr.wikipedia.org/wiki/Apprentissage_automatique?oldid=119068564
2. https://fr.wikipedia.org/wiki/Arbre_de_d%C3%A9cision
3. https://fr.wikipedia.org/wiki/M%C3%A9thode_des_k_plus_proches_voisins
4. https://fr.wikipedia.org/wiki/R%C3%A9seau_de_neurones_artificiels
5. https://fr.wikipedia.org/wiki/Mod%C3%A8le_de_m%C3%A9langes_gaussiens
6. https://fr.wikipedia.org/wiki/Analyse_discriminante_lin%C3%A9aire
7. https://fr.wikipedia.org/wiki/Machine_%C3%A0_vecteurs_de_support
8. https://fr.wikipedia.org/wiki/Apprentissage_automatique?oldid=119068564
15. https://fr.wikipedia.org/wiki/Algorithme_g%C3%A9n%C3%A9tique
16. http://biochimej.univangers.fr/Page2/COURS/8ModuleL1CSG/2ConfBioInformatique/3PresentHTML/1ConfBioInfoCSG.htm
17. http://www.rts.ch/decouverte/sciences-et-environnement/4637771
19. https://larlet.fr/david/biologeek/archives/20040930-la-bio-informatique-bioinfo-pour-les-intimes
20. http://www.fournier-majoie.org/fr/domaines-action/les-biomarqueurs-du-cancer
22. https://www.lri.fr/~antoine/Courses/DEA-I3/Tr-selection-attributs.pdf
26. http://decsai.ugr.es/~herrera/fl-ga.html
31. https://fr.wikipedia.org/wiki/K-moyennes
32. https://fr.wikipedia.org/wiki/Carte_auto_adaptative
35. http://www.fournier-majoie.org/fr/domaines-action/les-biomarqueurs-du-cancer