REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE FERHAT ABBAS-SETIF

MEMOIRE

Présenté à la Faculté des Sciences
Département d'informatique
Pour L'Obtention du Diplôme de

MAGISTER

Option : Ingénierie des Systèmes Informatiques

Par

Melle. GAGAOUA Meriem

THEME

Apprentissage et fouille de données par les algorithmes bio-inspirés : Application à la reconnaissance de caractères arabes manuscrits

Soutenu le : 21/01/2012

Devant le jury

Président :	**Dr. A.KHABABA**	**M.C Université Ferhat Abbas Sétif**
Rapporteur :	**Dr. A.MOUSSAOUI**	**M.C Université Ferhat Abbas Sétif**
Examinateur :	**Dr. O. KEZAR**	**M.C Université De Biskra**

Résumé

Dans ce travail nous avons traiter la problématique de la reconnaissance de l'écriture arabe manuscrite. Nous avons couplé deux algorithmes bio-inspirés, un système immunitaire artificiel pour la reconnaissance de mots et un algorithme génétique pour la sélection d'attributs de classification.

L'écriture arabe diffère des autres écritures par sa nature complexe à savoir la forte cursivité, la présence de points diacritiques, présence de mots ligaturés, elle a plusieurs fontes(Kofi, Thuluth et Diwani) et etc. A part les professionnels, d'autres scripteurs ne peuvent pas suivre toutes les règles de l'écriture manuscrite ce qui cause plus de difficulté pour la reconnaissance et rend la conception d'un OCR(optical caracter regognition) pour cette langue une véritable problématique.

Plusieurs travaux traitant l'écriture arabe ont été réalisés, que se soit imprimée ou manuscrite, en ligne ou hors ligne, suivant l'approche analytique ou holistique et en utilisant des techniques d'extraction de caractéristiques (structurelles, statistiques, basé sur les transformés et etc) et de reconnaissance (les techniques structurels, statistiques, basés les réseaux de neurones ou les chaînes de Markov...) différentes.

Pendant les derrières décennies l'inspiration des systèmes informatiques de la biologie a prouvé son succès dans la résolution des problèmes complexes ce qui a donné naissance aux algorithmes bioinspirés tel que les systèmes immunitaires artificiels, les algorithmes génétiques, les réseaux de neurones, les fourmis artificiels et etc.

Dans notre contribution nous avons adapté une nouvelle approche qui se base sur la méthode de fouille de données à savoir la classification ainsi que ses techniques qui s'inspirent de la biologie en l'occurrence les algorithmes bio-inspirés. Notre algorithme hybride immuno-genetic se base sur les systèmes immunitaires artificiels et les algorithmes génétiques. Pour la reconnaissance des mots arabe, nous avons utilisé les systèmes immunitaires artificiels qui ont la faculté de mémorisation et de bon apprentissage, de plus ils sont bien utilisés dans le domaine de reconnaissance de formes, pour l'optimisation et l'amélioration des résultats de classification nous procédons par la sélection d'attributs de classification, pour se faire nous avons choisi d'utiliser les algorithmes génétiques, qui sont réputés être de bons optimiseurs et bien appliqués dans ce genre d'application.

Nous avons organisé notre mémoire en quatre chapitres :

Dans **le chapitre 1**: nous avons présenté quelques généralités sur la fouille de données à savoir ses composants, ses méthodes ses techniques ainsi que ses domaines d'application. Nous avons détaillé l'une des méthodes de la fouille de données qui est la classification

car elle est la plus adaptée pour le cas de notre application (la reconnaissance de l'écriture arabe). Nous avons présenté aussi les méthodes de sélection d'attributs. La sélection d'attributs a pour but d'améliorer le taux de classification. En fin nous avons présenté quelques notions sur l'apprentissage artificiel qui est utilisé par certaines techniques de la fouille de données notamment les algorithmes bio-inspirés qui seront utilisés dans notre application et qui seront abordés dans le chapitre 3.

Dans le **chapitre 2** : nous avons présenté un état de l'art sur les OCR en générale et OCR arabe en particulier. Nous avons présenté les différents aspects de la reconnaissance de l'écriture, notre application se découle du type d'écrite manuscrite et du mode d'acquisition hors ligne. Puis nous avons montré l'organisation générale d'un OCR. Après nous avons détaillé la reconnaissance optique de l'écriture arabe, dans lequel nous avons présenté l'écriture arabe, ses caractéristiques, ainsi que les problèmes rencontrés dans sa reconnaissance particulièrement la manuscrite. Puis nous avons exposé le processus de reconnaissance commençant par la première étape le prétraitement qui engendre une série d'opérations(tel que la binarisation, le redressement, la squelétisation...), puis la segmentation qui est une tâche très délicate et très difficile où nous avons présenté quelques méthodes utilisés ainsi que les problèmes qu'elles rencontrent qui sont causés par la nature complexe de l'écriture arabe, vient après l'étape d'extraction de caractéristiques où nous trouvons les caractéristiques structurelles qui sont plus adaptés à l'écriture arabe manuscrite, les caractéristiques statistiques, basés sur les transformés ..., l'avant dernière étape qui est l'étape de reconnaissance où nous trouvons plusieurs méthodes (tel que les méthodes structurelles, statistiques, les réseaux de neurones ...) ainsi que quelques travaux réalisés et enfin l'étape de poste traitements qui a pour but d'améliorer le taux de reconnaissance. Nous avons présenté à la fin du chapitre un tableaux dans lequel nous avons montré quelques travaux réalisés ainsi que les méthodes utilisés dans le processus de reconnaissance.

Dans le **chapitre 3** :nous avons présenté quelques techniques de fouille de données qui s'inspirent de la biologie. Nous avons fait une étude sur quatre algorithmes bio-inspirés, à savoir les systèmes immunitaires artificiels, les réseaux de neurones, les algorithmes génétiques, les colonies de fourmies,... et une comparaison entre ces quatre algorithmes. Nous avons présenté leurs origines, caractéristiques, principes de fonctionnement, modèles ainsi que leurs domaines d'application. Nous avons choisi deux algorithmes : les systèmes immunitaires artificiels et les algorithmes génétiques pour les avantages qu'ils offrent par rapports aux autres, ainsi le premier pour faire la classification(la reconnaissance des noms des villes) et le deuxième pour la sélection de la liste d'attributs de classification.

Dans le **chapitre 4**: nous avons présenté notre contribution, notre approche bio-inspiré(algorithme hybride), les attributs de classification utilisées(caractéristiques struc-

turelles d'un mot) les outils de développement et à la fin nous avons discuté les résultats.

En fin nous avons présenté dans la conclusion quelques perspectives liées au domaine.

Notre approche a fait l'objet de communication présentée dans l'annexes A.

Table des matières

Table des figures

Liste des tableaux

Introduction Générale

L'une des applications prépondérantes et les plus intéressantes du domaine de la prise de décision est sans doute la fouille de données. Cette dernière consiste à extraire des connaissances à partir des données variés et hétérogènes. Pour se faire, elle accomplie des tâches tel que la recherche d'associations, le clustering, la classification et etc.

Le problème de la classification de données a été identifié comme une des problématiques majeures en fouille de données. La classification consiste à examiner les caractéristiques d'un objet et lui attribuer une classe. Cette classe est connue par la machine en utilisant des méthodes d'apprentissage artificiel.

L'apprentissage essaye de reproduire l'apprentissage naturel. C'est la science qui cherche et établie les liens entre les principes généraux d'apprenabilité et les méthodes et outils permettant de réaliser un apprentissage dans un contexte particulier. La classification utilise un apprentissage supervisé.

Il existe une multitude de techniques de classification qui ont des origines diverses et souvent multiples. Certaines sont issues des statistiques, d'autres proviennent des recherches en intelligence artificielle, certaines techniques s'inspirent de phénomènes biologiques ou de la théorie de l'évolution tel que les systèmes immunitaires artificiels, les réseaux de neurones ...etc.

Parmi les problèmes traités par la classification, la reconnaissance optique des caractères OCR (Optical Character Recognition) qui dérive du domaine de la reconnaissance de formes et qui occupe une place importante dans la recherche scientifique. Le manuscrit de la langue arabe est particulier et diffère de celui des langues latines d'où la conception d'un OCR pour cette langue est une véritable problématique. Plusieurs recherches se sont dirigées vers cet axe et plusieurs techniques de classification de données sont appliquées à la reconnaissance de l'écriture arabe: les réseaux de neurones artificiels, les modèles de Markove cachés(MMC), les algorithmes génétiques, la logique floue et etc. Ces méthodes ont leurs avantages mais révèlent une insuffisance envers la nature complexe de l'écriture

arabe (cursive, présence des points diacritiques, pseudo mots ligaturés et etc.).

Dans ce mémoire nous proposons une nouvelle approche de reconnaissance de l'écriture arabe manuscrite et cela en hybridant entre deux algorithmes bio-inspirés à savoir les systèmes immunitaires artificiels et les algorithmes génétiques. Nous avons utilisé un système immunitaire artificiel pour la reconnaissance des mots, puis nous avons appliqué une méthode d'optimisation avec les algorithmes génétiques pour sélectionner la combinaison d'attributs appropriés pour la classification immunitaire et qui génère le meilleur taux de classification.

Notre mémoire est organisé comme suit :

Dans le premier chapitre, nous présentons quelques généralités sur la fouille de données, la classification, la sélection d'attributs et l'apprentissage. Dans le deuxième nous présentons un état de l'art sur les OCR en générale et OCR arabe en particulier. Dans le troisième nous présentons une étude sur quatre algorithmes bio-inspirés, à savoir les systèmes immunitaires artificiels, les réseaux de neurones, les algorithmes génétiques, les colonies de fourmies, puis nous avons fait une comparaison entre ces quatre algorithmes et tirés deux pour les utiliser dans notre système. Dans le dernier chapitre nous présentons notre approche bio-inspiré, les attributs de classification utilisée, et à la fin une discutions des résultats. Nous terminons par une conclusion générale où nous exposons aussi les perspectives pour notre système.

Chapitre 1

Fouille de données

Depuis quelques années, une masse grandissante de données est générée de toute part et dans différents domaines. Les techniques usuelles analysant ces données sont insuffisantes d'où le besoin d'une nouvelle génération d'outils et de théories pour aider à extraire les informations utiles (les connaissances) à partir des volumes de données numériques qui croissent rapidement. Ces théories et outils sont le sujet d'un nouveau domaine appelé extraction de connaissances à partir de données dont le coeur est la fouille de données.

1.1 Extraction des connaissances à partir des données

La découverte de connaissances à partir de données KDD(Knowledge Discovery in Databases) est une modélisation et analyse exploratoire automatique d'un ensembles de données large. Le KDD est un processus organisé pour l'identification de configurations valides, nouvelles, et compréhensibles à partir d'un ensemble de données larges et complexe. La fouille de données est le coeur du processus du KDD. Elle utilise des algorithmes qui explorent les données, développent un modèle et découvrent des configurations inconnues. Le modèle est utilisé pour comprendre les phénomènes de l'analyse et de la prévision des données [1].

Le processus de découverte de connaissances est itératif et interactif il suit plusieurs étapes qui sont illustrées par le schéma de la figure 1.1, ce processus commence par la détermination des buts du KDD et se termine par l'implémentation des connaissances découvertes comme suit :

1. **Compréhension du domaine d'application et détermination des butes** : c'est la première phase préparatoire. Les buts de l'utilisateur final et l'environnement

de développement dans lequel se déroule le processus du KDD doivent être cernées, puis l'ensemble d'opérations de prétraitements sera lancé comme sera décrit dans les trois prochaines étapes.

2. **Sélection des données**: la sélection des données sur lesquels sera effectuée la découverte est primordiale. Cela inclue, la découverte des données valables, l'obtention des données nécessaires puis leur intégration dans un seul ensemble de données. Ce processus est très important car la fouille de données apprend et découvre à partir de données valables. Si quelques attributs importants manquent alors l'étude entière peut échouer.
3. **Prétraitement des données** : dans cette étape, la fiabilité des données est améliorée. Les données bruitées sont supprimées tel que les données doubles qui peuvent se révéler gênantes parce qu'elles vont donner plus d'importance aux valeurs répétées. Un contrôle sur les domaines des valeurs permet de retrouver des valeurs aberrantes. Un algorithmes de prédiction de données manquantes peut être mis en oeuvre.
4. **Transformation des données**: dans cette étape les méthodes incluent la réduction de dimension (telle que la sélection et l'extraction des caractéristiques) et la transformation des attributs (la discrétisation des attributs numériques et la transformation fonctionnelle). Cette étape est souvent cruciale pour le succès du projet de KDD mais elle est usuellement spécifique au projet. Cependant, si nous n'utilisons pas les bonnes transformations au début, nous pouvons obtenir un effet étonnant. Après avoir terminé les étapes précédentes , les prochaines étapes seront reliées à la fouille de données.
5. **Choix d'une tâche de fouille de données appropriée** : dans cette étape le choix d'une tâche de fouille de données est possible, tel que la classification, la régression, le clustering etc. Ceci dépend généralement des buts du KDD et également des étapes précédentes. Il y a deux buts importants dans la fouille de données: la prédiction et la description. La pluparts des méthodes de fouille de données sont basées sur l'apprentissage inductif, où le modèle est construit explicitement ou implicitement par la généralisation à partir d'un ensemble d'exemples d'apprentissage.
6. **Choix de l'algorithme de fouille de données**: cette étape inclue la sélection des méthodes spécifiques qui seront utilisées pour la recherche des connaissances.
7. **Application de l'algorithme de fouille de données** : finalement l'implémentation de l'algorithme de fouille de données est réalisée. Dans cette étape l'algorithme peut être utilisé plusieurs fois jusqu'à ce qu'un résultat satisfaisant soit obtenu.
8. **Evaluation**: dans cette étape l'évaluation et l'interprétation des configurations extraites avec le respect des buts définis dans la première étape auront lieu. Ici les étapes de prétraitements sont considérées ainsi que leurs effets sur les résultats des

algorithmes de fouilles de données. Cette étape se concentre sur la compréhensibilité et l'utilité du modèle induit. Ici la connaissance découverte est également documentée pour des utilisations futures.

9. **Utilisation des connaissances découvertes** : dans cette étape la connaissance est prête à être incorporée dans un autre système pour d'autre actions. En fait le succès de cette étape détermine l'efficacité du processus entier de KDD.

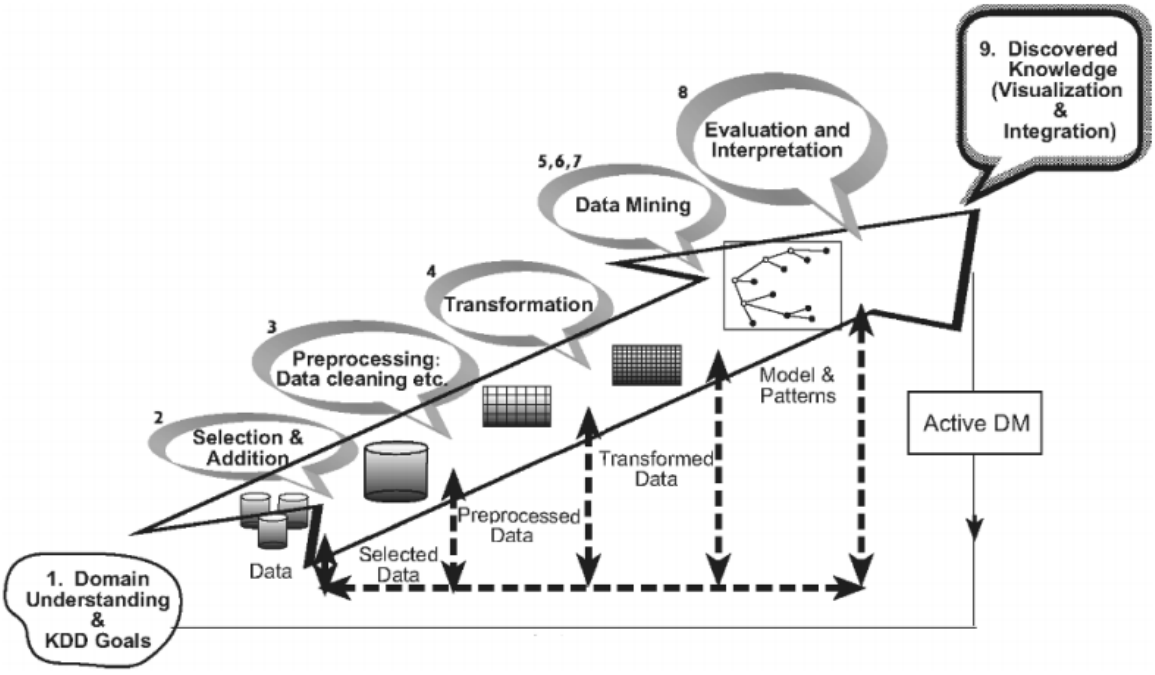


FIG. 1.1 – *Schéma d'un processus de fouille de données*
[1]

1.1.1 Méthodes de fouille de données

Il existe plusieurs méthodes de fouille de données utilisées pour différents buts. On distingue deux types: les méthodes orientées vérification et les méthodes orientées découverte comme illustré dans la figure 1.2. Les méthodes orientées découverte identifient automatiquement les configurations à partir des données. Elle consistent en méthodes de prédictions et de descriptions. Les méthodes descriptives sont orientées à l'interprétation de données tel que le clustering, la summarisation, la visualisation etc. Les méthodes prédictives visent à établir automatiquement un modèle comportemental qui obtient de nouveaux et invisibles échantillons et peut prévoir les valeurs d'une ou plusieurs variables liées à l'échantillon. Il développe également les configurations, qui forment la connaissance découverte d'une manière compréhensible et facile à utiliser. Parmi ces méthodes

on trouve celles qui se basent sur la régression et d'autres qui se basent sur la classification tel que les arbres de décision, les réseaux de neurones, SVM etc.

La plus parts des méthodes orientées découverte(en particulier quantitatives) sont basées sur l'apprentissage inductif où le modèle est construit explicitement ou implicitement par généralisation à partir d'un ensemble d'exemples. Les méthodes orientées vérification, procèdent en évaluant des hypothèses proposées par une source externe(expert, ...). Ces méthodes incluent les méthodes statistiques traditionnelles tel que le test d'hypothèses (t-test des moyennes), analyse de variance etc. Ces méthodes sont moins associées à la fouille de données par rapport aux méthodes orientées découverte, car la plus part des méthodes de fouille de données sont concernées par découvrir une hypothèse plutôt que tester celle qui sont déjà connues [1] .

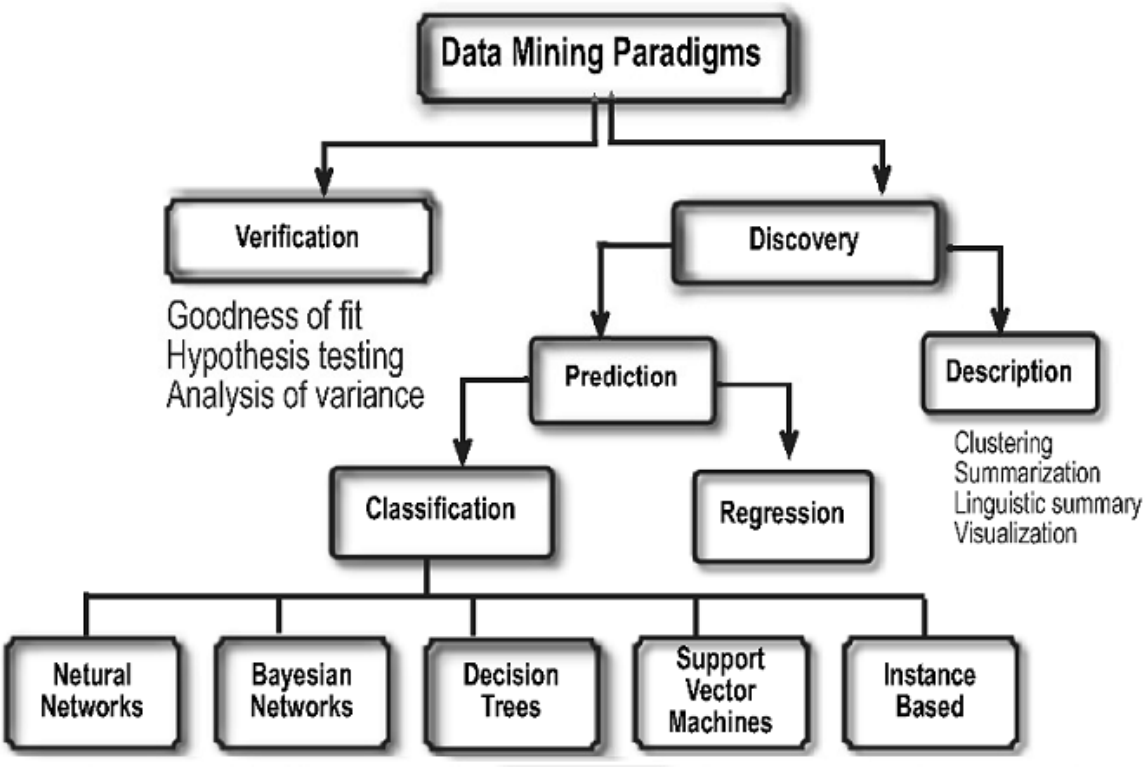


FIG. 1.2 – *Méthodes de fouilles de données*
[1]

1.1.2 Composantes des algorithmes de fouille de données

Trois composantes principales peuvent être identifiées dans un algorithme de fouille de données [6] :

1. **Modèle de représentation**: est le langage utilisé pour décrire les formes à dé-

couvrir. Si la représentation est trop limitée, alors le temps d'apprentissage est nul ou bien les exemples peuvent construire un modèle précis pour les données. Il est important qu'un analyste de données comprend entièrement les prétentions représentatives qui pourraient être inhérentes dans une méthode particulière. Il est également important qu'un concepteur d'algorithme formule clairement quelles prétentions représentatives sont faites par un algorithme particulier.

2. **Modèle d'évaluation**: les critères d'évaluation sont des rapports quantitatifs (ou une fonction de fitness) qui décrivent de combien un modèle particulier se rapproche du processus de découverte de connaissance.
3. **Méthode de recherche** : elle se compose de deux éléments les paramètres de recherche et le modèle de recherche.

1.1.3 Techniques de fouille de données

Il existe une variétés de techniques de data mining pour accomplir ses tâches tel que les réseaux de neurones , les arbres de décision, les algorithmes génétiques , les K plus proches voisins (KPP), les SVM, K- moyennes, les chaînes de markoves ...etc. Nous nous intéressons dans ce mémoire aux techniques utilisées dans la classification et inspirées de la biologie tel que les systèmes immunitaires artificiels, les réseaux de neurones ...etc.

1.1.4 Domaines d'application

Le succès des méthodes de data mining a intégré cette nouvelle science dans tout les domaines d'intelligence artificielle. Les méthodes de data mining sont applicables au problème d'estimation, prévention, analyse de risque, catégorisation, reconnaissance et etc. Nous citons ici quelques applications connues du data mining:

- **Multimédia:** les techniques du data mining sont appliquées aux bases de données multimédias pour la résolution de certains problèmes tel que la recherche d'image par le contenu, la reconnaissance de forme, la reconnaissance de la voie et etc.
- **Application de finance:** la fouille de données a fait ses premiers succès dans le domaine des finances. Des problèmes de prédiction et de prévention ont fait l'objet des méthodes statistiques qui ont donné naissance à des algorithmes spécialisés. Des méthodes pour l'analyse du comportement des clients, l'analyse des risques(assurances), l'estimation de rentabilité(banques) sont utilisées dans le secteur des finances.
- **Web mining:** le web mining est l'analyse des données du web par les techniques du data mining. On distingue différentes tâches du web mining: web content mining

(texte, image,...), Web structure mining (liens hypertextes,...) et Web usage mining (analyse des fichiers logs client et serveur).

- **Texte Mining:** la fouille de texte est utilisée dans divers domaines, les moteurs de recherche, la reconnaissance de l'écriture imprimée ou manuscrite et etc.
- **Médecine:** la fouille de données s'applique aussi dans le domaine médicale, tel que le diagnostic automatique ou l'aide au diagnostic (découverte de la maladie du patient d'après ses symptômes), recherche du médicament le plus approprié à une maladie et etc.

1.2 Classification de données

La classification de données est l'une des tâches importantes du datamining. Elle permet de prédire si une instance de donnée est membre d'un groupe ou d'une classe prédéfinie. Les classes sont des groupes d'instances avec des profils particuliers [2]. Pour ce faire les classe doivent être connus à l'avance.

1.2.1 Le processus de classification

Le processus de classification se déroule en deux étapes principales qui sont illustrés dans le schéma de la figure 1.3 :

1. **Construction du modèle**
   La construction du modèle de classification se fait à partir d'un ensemble d'apprentissage (training set). Chaque instance est supposée appartenir à une classe prédéfinie où la classe de cette instance est déterminée par l'attribut classe, l'ensemble des instances d'apprentissage est utilisé dans la construction du modèle. Le modèle est représenté par des règles de classification, arbres de décision, formules mathématiques, des réseaux de neurones... etc.
2. **Utilisation du modèle**
   Le modèle établie dans la phase précédente sera utilisé dans la classification de nouvelles instances. Et le taux d'erreur de ce modèle sera calculé pour tester ses performances.

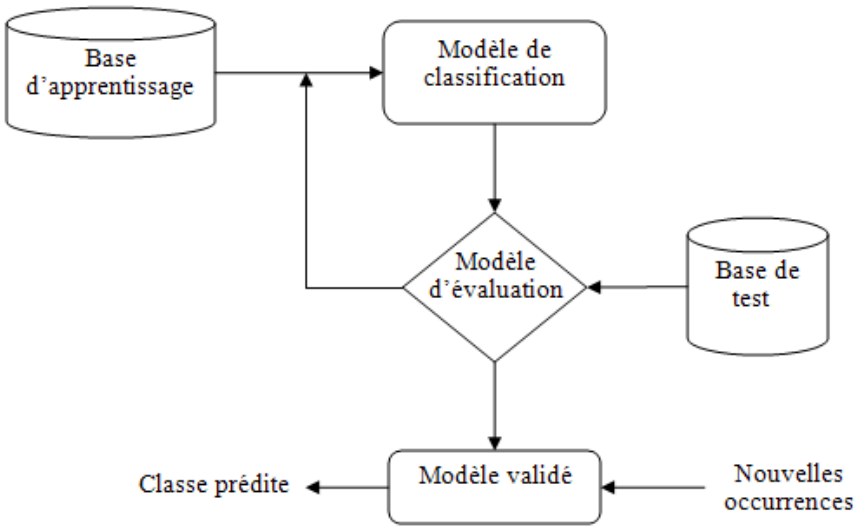


FIG. 1.3 – *Processus de classification*

1.2.2 Evaluation des méthodes de classification

La résolution d'un problème de classification s'effectue en comparant des modèles afin de choisir le plus apte à résoudre le problème posé. L'évaluation des modèles est donc un préalable inévitable à la classification. Elle est nécessaire pour connaître les performances d'un modèle et déterminer s'il est globalement significatif. Dès lors, deux objectifs se dégagent: l'évaluation et la comparaison de modèles en vue de la sélection. La sélection du modèle idéal peut être envisagée :

- En comparant différentes méthodes de classification pour un même sous-ensemble de variables ;
- En comparant différentes méthodes de sélection de variables, pour une même méthode de classification ;
- En comparant simultanément des méthodes de classification et des méthodes de sélection de variables.

1.2.2.1 Mesures de la qualité d'un modèle

La qualité d'un modèle de classification est souvent définie par son taux de classification, indiquant sa capacité et ses performances de discrimination, obtenue par le taux d'observations bien ou mal classées (erreur de classification). Cependant, d'autres éléments moins quantitatifs, peuvent contribuer à rendre un modèle intéressant [4], comme :

- **La robustesse**: qui donne une information sur le pouvoir de généralisation du modèle, donc sa sensibilité aux variations des observations. En d'autres termes, le

modèle doit être le moins dépendant possible des observations d'apprentissage afin d'éviter le phénomène de sur-apprentissage. En outre, un modèle robuste est forcément dépendant des variables d'entrée qu'il observe. Par conséquent, la légitimité de la présence de ces variables doit être vérifiée et celles-ci doivent pouvoir être recueillies sans difficulté. Dans le cas contraire, le modèle doit pouvoir s'adapter à la présence de valeurs manquantes. Selon [4], un modèle robuste doit pouvoir s'adapter également aux variables qui évolueraient dans le temps. Néanmoins, il précise légitimement que cette durée doit être raisonnable.

- **La parcimonie**: suggère de réaliser un modèle le plus simple possible. La simplicité du modèle facilite sa compréhension et sa lisibilité, par exemple la lecture de la relation liant ses entrées à ses sorties ;
- **Le coût calculatoire:** peut être un paramètre important dans l'évaluation d'un modèle, selon l'utilisation que l'on souhaite faire de ce dernier. En effet, si l'apprentissage du modèle doit être réalisé en ligne, de manière à procéder à des ajustements en situation, il peut être souhaitable que l'apprentissage puisse être réalisé relativement rapidement. D'autre part, même en utilisation hors ligne, plus l'apprentissage est rapide, plus le nombre de tests et d'ajustements peut être effectué afin d'affiner le modèle.

Ces trois éléments montrent que la qualité d'un modèle est fortement liée à son pouvoir de généralisation. Plus sommairement, Webb dans [5] décompose l'évaluation de la performance des modèles de classification en deux facteurs : le pouvoir discriminant de sa règle de classification et sa fiabilité.

Selon Han [8], une bonne méthode de classification automatique, dans le domaine de la fouille de données, doit :

- Traiter de grandes bases de données, aussi bien au niveau du nombre d'objets que du nombre de variables,
- Traiter des données de tout type (numériques, binaires, catégorielles, ordinales, etc.).
- Découvrir des classes de profil arbitraire.
- Requérir très peu de connaissance du domaine d'application.
- Gérer des données bruitées (outliers, données manquantes, données erronées etc.).
- Etre paramétrable pour pouvoir satisfaire certaines contraintes.
- Produire des résultats simples, clairs, et utilisables directement.

1.2.3 Techniques de classification

Il existe une multitude de techniques de classification qui ont des origines diverses et souvent multiples. Certainnes sont issues des statistiques tel que les KPP, classification bayésienne..., d'autres proviennent des recherches en intelligence artificielle tel que les arbres de décisions , les règles d'induction... , certainnes s'inspirent de phénomènes biologiques ou de la théorie de l'évolution tel que les systèmes immunitaires artificiels, les réseaux de neurones, les algorithmes génétiques ...etc. Dans notre mémoire nous nous intéressons aux algorithmes bio-inspirés pour faire la classification (reconnaissance) qui seront détaillés dans les chapitres qui suivent.

1.2.4 Sélection d'attributs

La qualité des résultats de classification dépend fortement du type de classifieur et des attributs de classification utilisés. Pour l'amélioration des résultats de classification et/ou la réduction du temps de traitement , l'idée est d'utiliser des méthodes de sélection d'attributs, qui sélectionnent parmi les attributs originales, les plus pertinents de manière à former un sous-ensemble d'attributs préservant l'information utile.

L'organigramme de la figure 1.4 proposé par [9], illustre la procédure générale de la sélection d'un sous-ensemble de variables. Cette procédure repose sur deux éléments principaux : le critère d'évaluation et la procédure de recherche.

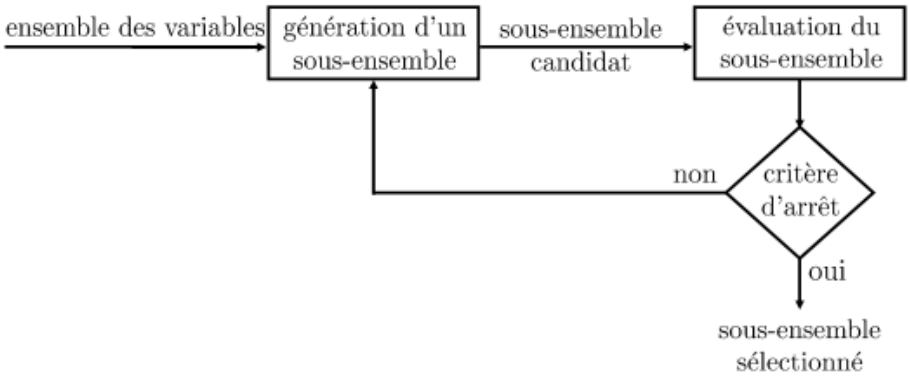


FIG. 1.4 – *Procédure de recherche d'un sous-ensemble de variables* [9]

1.2.4.1 Approches de sélection

Pour la sélection d'un sous ensembles d'attributs deux approche principale se distinguent, wraper et filter. Elles se différencient en fonction de la participation de l'al-

gorithme d'apprentissage dans la sélection du sous-ensemble de variables. La figure 1.5 illustre ces deux approches.

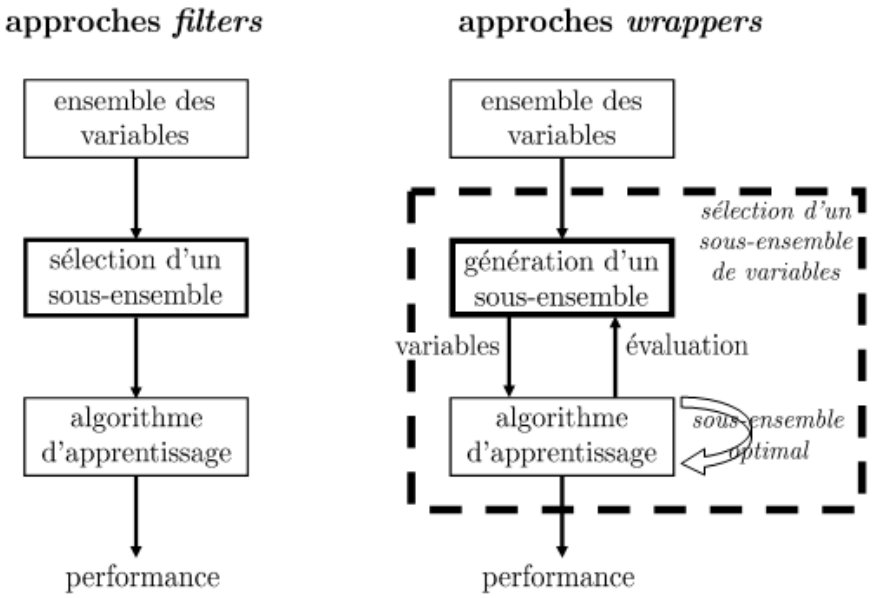


FIG. 1.5 – *Approches principales de sélection d'attributs*

1. **Approche enveloppante (*wraper*)**

   Cette approche utilise un classifieur comme méthode d'évaluation. Bichop [10] suggère d'utiliser des méthodes de classification plus simples et plus rapides, telles que des techniques linéaires, pour sélectionner les variables et ainsi, générer le modèle final avec des méthodes de classification sophistiquées à partir du sous-ensemble préalablement déterminé. Cependant, par cette approche, le sous-ensemble de variables, obtenu durant la sélection, ne sera pas forcément optimal pour la conception du modèle. En effet, Liu et Yu dans [11] soulignent judicieusement qu'un sous-ensemble de variables peut être optimal suivant un certain critère et peut ne plus l'être pour un autre. Ils notent ainsi l'importance et l'influence du critère d'évaluation dans le processus de sélection. Le critère d'arrêt utilisé dans cette approche est le taux de reconnaissance.

2. **Approche filtrante (*filter*)**

   Cette approche permet de rechercher des sous-ensembles de variables, sans utiliser un algorithme d'apprentissage [12]. En effet, en présence d'observations étiquetées, le choix d'un sous-ensemble de variables peut se faire en considérant l'habilité du sous-ensemble à discriminer les classes. Dans ce cas, la pertinence d'une variable pourrait être définit par une mesure de séparabilité des classes, ou encore, par une évaluation du recouvrement entre les classes [13]. Cette pertinence s'obtiendrait indépendamment d'un algorithme d'apprentissage.

3. **Approche hybride**,
   Cette approche tente de tirer avantage des précédentes approches, en exploitant leurs différents critères d'évaluation dans plusieurs étapes de la recherche du sous-ensemble [14]. Cette dernière approche peut être privilégiée en présence d'un nombre de variables très important.

En présence d'un nombre de variables très important. Dans la pratique, on pourrait dans une première étape faire une pré-sélection par des méthodes de type filter afin de réduire le nombre de variables. Puis, pour optimiser la sélection, une deuxième étape fondée sur une approche de type wrapper pourrait être réalisée afin d'obtenir la sélection du sous-ensemble final.

1.2.4.2 Critères d'évaluation des attributs

Pour la sélection de variables il est nécessaire de définir une mesure de pertinence, faisant état de la qualité de la variable ou du sous-ensemble de variables sélectionnées. Bennani dans [15] définit une variable pertinente telle que sa suppression entraîne une détérioration des performances du système de classification.

Idéalement, en classification supervisée, le critère d'évaluation d'un sous-ensemble de variables pourrait être fondé sur le taux de classification. Ce dernier serait obtenu par l'évaluation des performances de généralisation du modèle, une fois l'apprentissage réalisé ; les entrées de ce modèle seraient composées des variables pré-sélectionnées. Ceci est addapté avec l'approche wraper.

Avec les approches de type filter, Webb et al [5,13] décrivent rigoureusement plusieurs mesures d'évaluation, afin d'estimer la capacité d'une variable ou d'un sous-ensemble à séparer les classes qui peuvent être regrouper en plusieurs catégories tel que la mesure des distances probabilistes(la plus connu de mahalanobis), la mesure de dépendance et etc.

L'intérêt des méthodes d'évaluation de type filter réside dans leur rapidité d'exécution, grâce notamment à la non-utilisation d'outils de classification. Cependant, comme souligné par Liu et Yu [11], le sous-ensemble optimal pourrait se révéler inefficace une fois appliqué à un outil de classification. Cet inconvénient n'apparaît pas dans les méthodes de type wrapper, où l'évaluation est fondée directement sur l'outil de classification en contrepartie, cette démarche rend les méthodes wrapper fortement dépendantes du classifieur utilisé.

1.2.4.3 Méthodes de recherche

Il existe plusieurs techniques de recherche de sous ensembles d'attributs pertinents, ainsi suivant l'approche de sélection suivi nous distinguons les méthodes suivantes:

1. **Classement des variables**

   L'obtention d'un classement, indiquant la qualité de discrimination des variables, peut être réalisé indépendamment d'un outil de classification. Sans classifieur, il est donc nécessaire de définir un "'score"' pour évaluer la pertinence de chaque variable. L'évaluation des variables peut se faire par des approches de type *filter*, en employant un critère mesurant la séparabilité, comme FDR(*Fisher discriminant Ratio*) par exemple. Les variables sont ensuite triées afin d'obtenir un classement allant de la variable la plus pertinente à la moins pertinente. Ainsi, la sélection de q variables correspondrait aux q meilleures valeurs de "'scores"'.

2. **Sélection d'un sous-ensemble de variables**

   Les procédures de recherche s'emploient à générer des sous-ensembles dans l'espace des variables. La présence de p variables entraîne un nombre de p^2 combinaisons possibles. La recherche exhaustive de toutes ces combinaisons peut entraîner une explosion combinatoire pour remédier à ça, il est nécessaire de changer de stratégie de recherche, en se basant sur des procédures dites sous-optimales qui s'occupent de gérer la sélection et l'élimination des variables.

   Il existe plusieurs stratégies de de sélection d'attributs, suivant la classification de Liu et al dans [9] où ils considèrent les trois catégories suivantes : exhaustive, heuristique et non déterministe.

   Après la décision de la stratégie à adopter, le choix de la direction de recherche à prendre aura lieu. On dénombre trois types de stratégies :

   - **Stratégie ascendante**: qui commence à partir d'un ensemble vide de variables, puis ajoute progressivement des variables;
   - **Stratégie descendante**: qui commence avec toutes les variables, puis élimine progressivement les variable ;
   - **Stratégie aléatoire**: qui choisit aléatoirement un sous-ensemble de variables, puis ajoute ou retire progressivement des variables.

3. **Approches heuristiques**

   Les approches heuristiques répondent à une contrainte calculatoire, notamment lorsque le nombre de variables empêche une évaluation exhaustive. Ces approches permettent alors de faire un compromis entre le nombre de combinaisons à évaluer

et le coût global de l'évaluation. Les algorithmes opèrent par recherche séquentielle. Cette stratégie réduit le nombre de combinaisons à évaluer, en appliquant des recherches locales suivant une direction qu'elle soit : ascendante, descendante ou encore aléatoire. Parmi les méthodes utilisant une évaluation de type wrapper, les techniques heuristiques les plus connues sont fondées sur des sélections séquentielles ascendante et descendante, respectivement nommées sequential forward selection (SFS) et sequential backward selection (SBS). Ces méthodes identifient le meilleur sous-ensemble de variables en ajoutant ou en éliminant progressivement des variables.

4. **Approches déterministes**

   Ces méthodes ne pourront jamais s'extraire d'optimums locaux, même lors d'exécutions répétées de plusieurs processus. Une solution possible pour remédier à ce problème était par exemple de sélectionner aléatoirement un sous-ensemble de variables de départ. Ainsi, ce type d'approches serait à mi-chemin entre les heuristiques précédentes et les approches non déterministes. Dans cette perspective, nous pouvons évoquer la méthode RGSS(random generation plus sequential selection).

5. **Approches non déterministes: les algorithmes génétiques**

   Par opposition aux approches heuristiques, qui donnaient des résultats similaires à chaque exécution, des approches non déterministes, telles que les métaheuristiques, produisent des résultats différents à chaque exécution. Dérivées des heuristiques qui s'appuient sur la connaissance du problème pour trouver une solution, les métaheuristiques présentent l'avantage d'une certaine indépendance par rapport au problème à résoudre . Cet atout rend les métaheuristiques adaptables à bon nombre de problèmes parmi lesquels la sélection d'attributs. On peut citer parmi ces méthodes de cette approche le recuit simulé, la recherche tabou ou encore les algorithmes génétiques.

   Dans la sélection de variables, les algorithmes génétiques surpassent les méthodes heuristiques grâce à une plus grande diversité dans l'exploration de l'espace de recherche. Ceci fait des algorithmes génétiques, et plus généralement des approches non déterministes, des méthodes d'optimisation globale. Elles peuvent être adaptées lorsque l'espace de recherche devient trop vaste [16].

Compte tenu du nombre de combinaisons possibles dans le choix de l'approche à employer, entre le critère d'évaluation et la procédure de recherche, on aperçoit rapidement la difficulté pour choisir le processus de sélection. On a vu que le choix du critère d'évaluation pouvait dépendre du nombre de variables composant l'ensemble de départ. Plus ce nombre serait grand, plus on aurait tendance à choisir une méthode d'évaluation de

type filter. L'utilisation d'outils de classification pour les méthodes de type wrapper rend la sélection plus performante (du point de vue du classifieur), mais la rend dépendante du classifieur utilisé lors de la sélection. Ce point donne dans un contexte où l'outil de classification st susceptible de changer, un avantage indéniable aux méthodes filter.

1.3 Apprentissage Artificiel

L'apprentissage artificiel est la science qui cherche et établie les liens entre les principes généraux d'apprenabilité et les méthodes et outils permettant de réaliser un apprentissage dans un contexte particulier. Il essaye de mimer l'apprentissage naturel [7]. Il est , définit aussi par Cornuéjols et al dans [17], comme une notion englobant toute méthode permettant de construire un modèle réel à partir d'un ensemble de données soit en améliorant un modèle partiel (ou moins général), soit en créant complètement le modèle. La popularité croissante de l'apprentissage artificiel est certainement dû à son approche multidisciplinaire. En effet, de part la diversité des outils produits et des problèmes traités, l'apprentissage artificiel se trouve au carrefour de nombreuses disciplines. La figure 2.1 montre une liste non exhaustive de domaines avec lesquels il s'est exposé.

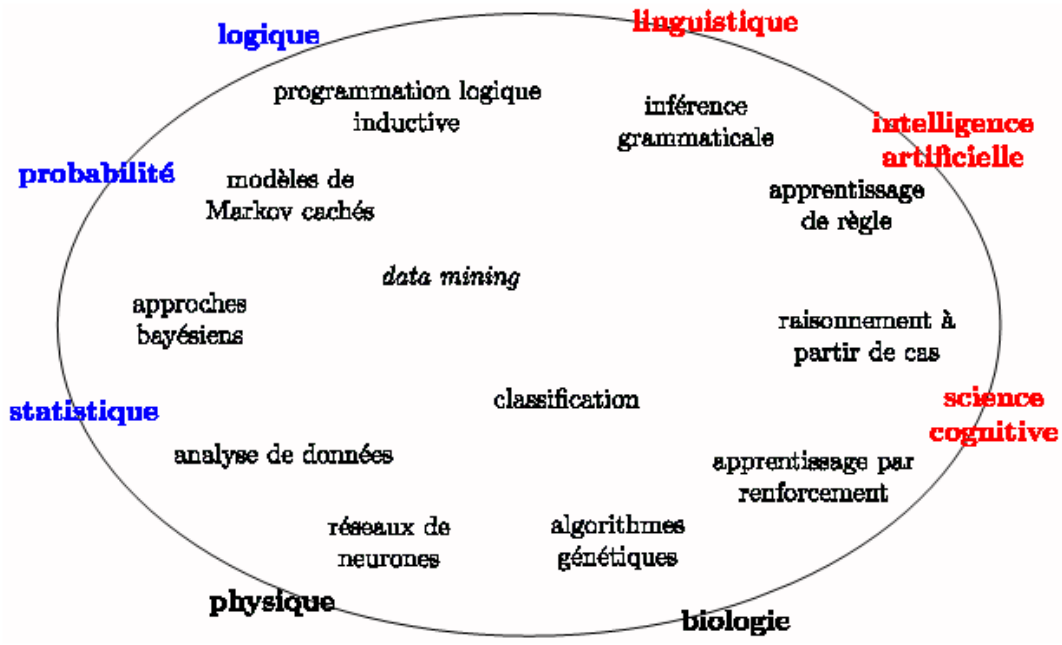


FIG. 1.6 – *Illustration des domaines scientifiques apparentés à l'apprentissage artificiel*

1.3.1 Techniques d'apprentissage

Un algorithme d'apprentissage reçoit un ensemble d'exemples d'apprentissage et doit produire des règles générales qui représentent les informations obtenues à partir de ces exemples. Il existe différents contextes d'apprentissage, à savoir, l'apprentissage supervisé, l'apprentissage non-supervisé, et l'apprentissage semi-supervisé (par renforcement).

- **Apprentissage supervisé**: est un apprentissage qui nécessite la présence d'un superviseur qui possède une connaissance approfondie de l'environnement dans lequel évolue le système. Dans ce cadre, disposant d'un ensemble des données préalablement étiquetées sous la forme d'entrées/sorties (les exemples d'apprentissage), le but est de chercher à trouver ou approximer la fonction qui permet d'effectuer automatiquement l'étiquetage le plus vraisemblable sur d'autres données (dites les exemples de test). Ce mode d'apprentissage peut servir à des fins d'analyse, de prise de décision et de prévision. Toutes les méthodes de classification sont donc des algorithmes d'apprentissage supervisés.
- **Apprentissage non supervisé**: il est caractérisé par l'absence complète de superviseur, il s'effectue sans interversion externe. L'apprentissage non-supervisé utilise des données sans étiquettes. Dans de telles conditions, l'apprenant ne reçoit aucune information indiquant quelles devraient être ses sorties ou même si celles-ci sont correctes. Il doit donc découvrir la structure des données à partir des corrélations existantes entre les exemples d'apprentissage qu'il observe. Ce mode d'apprentissage concerne plutôt des tâches d'analyse exploratoire des données. Toutes les méthodes de clustering sont des algorithmes d'apprentissage non supervisés.
- **Apprentissage par renforcement**: l'apprentissage par renforcement ou semi-supervisé fait référence à une classe de problèmes d'apprentissage, dont le but est d'apprendre, à partir d'expériences. Dans ce cadre l'apprenant ne dispose que d'indications imprécises sur la justesse de sa sortie (par exemple,échec/succès). Il s'agit donc de produire de plus en plus de sorties correctes en recourant à un processus d'essais et d'erreurs.

Les algorithmes d'apprentissage requièrent typiquement un ensemble d'exemples à partir duquel un modèle est construit, on parle alors d'un ensemble d'apprentissage. Dans un cadre supervisé, il est aussi indispensable d'avoir à disposition un autre ensemble d'exemples pour évaluer la validité de la solution trouvée par ce modèle, on parle alors d'un ensemble de validation ou de généralisation. Ces ensembles sont nécessairement disjoints et souvent établis à partir d'un ensemble unique d'exemples en le divisant en n parties égales et en construisant n modèles, en écartant à chaque fois une des n parties qui servira pour le

test et en utilisant les n - 1 autres parties pour l'apprentissage. Cette procédure est connue sous le nom de validation croisée (en anglais "*cross-validation*"). Elle permet notamment d'augmenter la capacité de généralisation d'un modèle d'apprentissage et d'arrêter le processus d'apprentissage avant de trop se spécialiser sur l'ensemble d'apprentissage, ce phénomène est connu sous le nom de "sur-apprentissage", ou "apprentissage par coeur" (ou encore "*overfitting*" en anglais). Dans la pratique, ce sont les capacités de généralisation d'un modèle qui vont établir les possibilités de l'appliquer à d'autres exemples que ceux vus au cours de la phase d'apprentissage.

1.4 Conclusion

La fouille de données est une branche très importante, elle offre plusieurs méthodes et techniques pour la résolution de différents problèmes. Parmi les domaines d'application de la fouille de données la reconnaissance de forme et en particulier la reconnaissance de l'écriture manuscrite arabe qui sera détaillée dans le prochain chapitre. La fouille de données offre plusieurs techniques pour la résolution des problèmes nous nous intéressons aux techniques qui s'inspirent de la biologie notamment les algorithmes bio-inspirés qui seront détaillés dans les chapitres qui suivent.

Chapitre 2

Reconnaissance de l'écriture Arabe

Toute information écrite peut être reprise et informatisée à différentes fins et dans plusieurs domaines (bureautique, reconnaissance de montants littéraux de chèques bancaires,tri du courrier dans les postes ...) d'où on trouve la reconnaissance optique des caractères (OCR) qui dérive du domaine de la reconnaissance de forme et qui occupe une place importante dans la recherche scientifique.

Contrairement au latin la reconnaissance de l'écriture arabe imprimée ou manuscrite reste encore aujourd'hui un domaine non complètement exploré. A. nazif [18] fus le premier à travailler sur la reconnaissance de l'écriture arabe dans sa thèse de master en 1975.

2.1 Les différents aspects de la reconnaissance optique de l'écriture

Un système de reconnaissance de caractère (OCR) est un programme désigné pour convertir des documents scannés vues par l'ordinateur comme des images vers des documents texte.

Les OCR différent les un des autres suivant le mode d'acquisition des données (" en-ligne " ou " hors-ligne "), le type d'écriture traitée (imprimée ou manuscrite), selon que l'analyse s'opère sur la totalité du mot ou par segments composés d'un seul caractère (les approches globales ou analytiques).

2.1.1 Le Type d'écriture

L'écriture peut être imprimé ou manuscrite:

- **Imprimée**: les caractères imprimés sont dans le cas générale alignés horizontale-

ment et séparés verticalement, ce qui simplifie la phase de lecture. La forme des caractères est définie par un style calligraphique (fonte) qui constitue un modèle pour l'identification [19].

- **Manuscrite**: dans le cas du manuscrit, les caractères sont souvent ligaturés et leurs graphismes sont inégalement proportionnés, provenant de la variabilité intra et inter scripteurs. Ce qui nécessite généralement l'emploi de techniques de délimitation spécifiques et souvent des connaissances contextuelles pour guider la lecture [20].

2.1.2 Le mode d'acquisition des données

L"acquisition des données peut être faite en ligne ou hors ligne:

- **La reconnaissance en ligne** : ce mode de reconnaissance s'opère en temps réel (pendant l'écriture) les caractères sont reconnus au fur et à mesure de leurs écriture à la main. Ce mode est réservé généralement à l'écriture manuscrite, c'est une approche "signal" où la reconnaissance est effectuée sur des données à une dimension. L'écriture est représentée comme un ensemble de points dont les coordonnées sont en fonction du temps. L'acquisition de l'écrit est généralement assurée par une tablette graphique munie d'un stylo électronique [21]
- **La reconnaissance hors ligne** : elle convient aux documents imprimés et les manuscrits déjà rédigés. Les données en entrée sont acquises via un scanner. Elle a comme tâche de déterminer quelles lettres ou mots sont présents dans l'image du texte scanné [22].

2.1.3 Approches de reconnaissance

La reconnaissance peut être faite sur la totalité du mot suivant l'approche holistique ou suivant l'approche analytique:

- **Approche globale**: elle se base sur une description unique de l'image du mot(vu comme une entité indivisible) et tente de le reconnaître en utilisant les caractéristiques du mot entier. Disposant de beaucoup d'informations (un vocabulaire large) rend la discrimination de mots proches très difficile et l'apprentissage des modèles nécessitant une grande quantité d'échantillons qui est souvent difficile à réunir pour l'apprentissage d'un classeur holistique. Cette approche reste parfaitement envisageable pour les vocabulaires réduits et distincts (exemple: reconnaissance de montants littéraux de chèques bancaires) [23].

- **Approche analytique**: cette approche traite le mot comme une collection de sous unités simples comme les caractères et procède en segmentant ces mots en unités, par exemple le mot: مريم est segmenté en 4 unités : م ـريـ ـيـ ـم .
  Cette approche est la seule applicable dans le cas de grands vocabulaires. Elle présente cependant des problèmes, surtout dans la segmentation de l'écriture manuscrite tel que : l'ambigüité dans la détermination des points de segmentation et la détermination de l'identité de chaque segment à cause de la variabilité de la forme des segments. [23]

2.2 L'organisation générale d'un système de reconnaissance de l'écriture

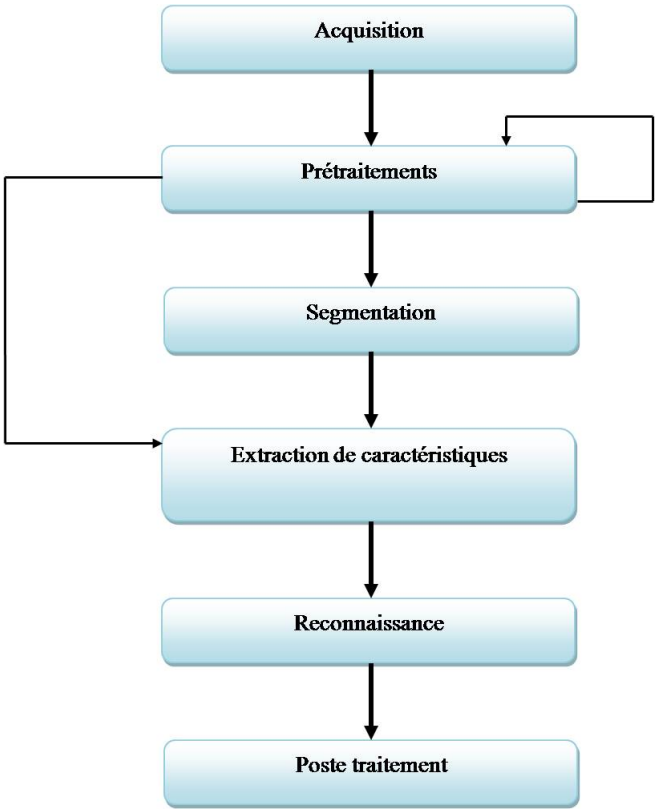


FIG. 2.1 – *Organisation générale d'un système de reconnaissance d'écriture*

La figure 2.1, représente un schéma générale d'un système de reconnaissance de l'écriture. Pour assurer une bonne reconnaissance, après l'acquisition des données, ces dernières passent par un ensemble d'opérations de prétraitement, puis une étape de segmentation aura lieu. Dans le cas de reconnaissance de mot en utilisant l'approche holistique le système passe directement à l'étape d'extraction de caractéristiques. Ces dernières seront

utilisées dans l'étape de classification, puis à la fin une étape d'amélioration des résultats pourra avoir lieu.

- **Acquisition:** l'acquisition des données peut être faite par une tablette en cas de reconnaissance en ligne (temps réel) où la numérisation des scripts dépend de la résolution de la tablette et de la vitesse d'échantillonnage. Dans le cas de la reconnaissance hors ligne, l'acquisition est effectué par un scanner où la qualité des données acquises dépend de la résolution de ce dernier.
- **Prétraitement :** c'est l'ensemble d'opérations qui apportent des transformations à une image dans le but de réduire le bruit superposé aux données. Comme opérations de prétraitements, nous pouvons citer le lissage, le redressement, etc.
- **La segmentation :** dans cette phase les différentes parties logiques de l'image sont séparées en blocs graphiques et blocs de texte, puis à partir des blocs de texte on extrait les lignes, puis les mots, ensuite les caractères .
- **Extraction de caractéristiques :** cette phase analyse les segments de textes et extrait un ensemble de caractéristiques qui servent à son identification.
- **La classification et reconnaissance :** c'est la phase de prise de décision dans un système OCR, elle utilise les caractéristiques extraites dans la phase précédente pour identifier des segments de texte suivant certaines règles.Cette phase peut utiliser des modèles de référence obtenus lors d'une phase d'apprentissage pour classer les données.
- **Poste traitement:** c'est la phase finale, elle améliore la reconnaissance par le raffinement des décisions prises dans la phase précédente, elle reconnaît les mots en utilisant le contexte. Elle est souvent implémentée comme un ensemble d'outils relatifs à la fréquence d'apparition des caractères, le lexique et à d'autres informations contextuels. [21]

2.3 Reconnaissance optique de l'écriture arabe

La reconnaissance de l'écriture arabe est un cas particulier des OCR.

2.3.1 Description et caractéristiques de l'écriture arabe

La langue arabe est une langue universelle, c'est la langue officielle de 25 pays de plus de 250 millions de population [32].

Les caractères de la langue arabe sont utilisés dans d'autres langues telles que farsi, jawi, etc. Les musulmans peuvent lire et écrire l'arabe car c'est la langue du Coran, le

No	Name	Isolated	Connected		
			Beginning	Middle	End
1	**Alif**	ا	ـا	ـا	ـا
2	**Baa**	ب	بـ	ـبـ	ـب
3	**Taa**	ت	تـ	ـتـ	ـت
4	**Thaa**	ث	ثـ	ـثـ	ـث
5	**Jeem**	ج	جـ	ـجـ	ـج
6	**Haa**	ح	حـ	ـحـ	ـح
7	**Khaa**	خ	خـ	ـخـ	ـخ
8	**Daal**	د	د	ـد	ـد
9	**Thaal**	ذ	ذ	ـذ	ـذ
10	**Raa**	ر	ر	ـر	ـر
11	**Zaay**	ز	ز	ـز	ـز
12	**Seen**	س	سـ	ـسـ	ـس
13	**Sheen**	ش	شـ	ـشـ	ـش
14	**Saad**	ص	صـ	ـصـ	ـص
15	**Shaad**	ض	ضـ	ـضـ	ـض
16	**Ttaa**	ط	طـ	ـطـ	ـط
17	**Dthaa**	ظ	ظـ	ـظـ	ـظ
18	**Ain**	ع	عـ	ـعـ	ـع
19	**Ghen**	غ	غـ	ـغـ	ـغ
20	**Faa**	ف	فـ	ـفـ	ـف
21	**Qaf**	ق	قـ	ـقـ	ـق
22	**Kaf**	ك	كـ	ـكـ	ـك
23	**Lam**	ل	لـ	ـلـ	ـل
24	**Mem**	م	مـ	ـمـ	ـم
25	**Noon**	ن	نـ	ـنـ	ـن
26	**Haa**	ه	هـ	ـهـ	ـه
27	**Wow**	و	و	ـو	ـو
28	**Yaa**	ي	يـ	ـيـ	ـي

TAB. 2.1 – *L'alphabet arabe et les différentes formes*

[24]

saint livre des musulmans. L'alphabet arabe et les différentes formes des caractères dans le mot sont décrits dans le tableau 2.1.

2.3.1.1 Les caractéristiques de l'écriture arabe

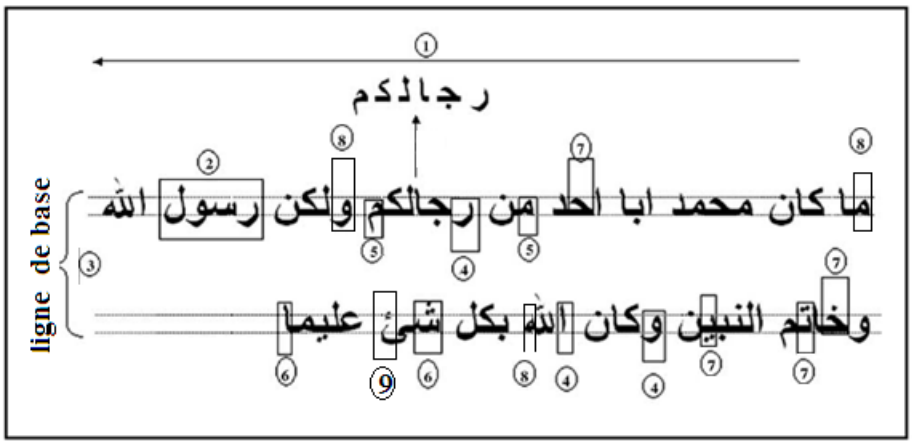


FIG. 2.2 – *Caractéristiques de l'écriture arabe*
[33]

Les caractéristiques de la langue arabe peuvent être résumées en :

1. L'arabe s'écrit de droite à gauche.
2. Un mot peut être constitué de deux ou plusieurs sous mots.
3. Les mots sont écrits suivant une ligne imaginaire appelée ligne de base.
4. L'arabe est toujours écrit cursivement et les mots sont séparés par des espaces , sauf avec six caractères qui ne peuvent être connectés qu'a droite ز ، و ، ا ، د ، ذ ، ر
   et lorsque ils apparaissent dans le mot il sera coupé en deux ou en plusieurs sous mots séparés par des petits espaces
5. La forme du caractère change selon sa position dans le mot où chaque caractère a deux ou quatre formes, ce qui augmente le nombre de classe à reconnaître de 28 à 84.
6. La taille des caractères diffère de l'un à l'autre et dans le même caractère selon sa position dans le mot.
7. Quinze caractères ont des points qui peuvent être au dessous ou au dessus de la partie principale. Quelques caractères partagent la même partie principale sauf les points pouvant faire la différence entre eux.

8. Quelques caractères contiennent des boucles.Une boucle est une caractéristique importante pour décrire un mot. Le caractère ڈ a une autre forme avec deux boucles ﮓ .
9. Le caractère Hamza (ء) , n'est pas vraiment une lettre, c'est une forme complémentaire qui est utilisée avec la lettre ك dans la forme isolée ou à la fin et elle peut être utilisée isolée ou avec les lettres ا ، و ، ى
10. Elle utilise des diacritiques présentés dans le tableau 4.3 qui sont utiles, des fois, pour montrer le vrai sens d'un mot et sa prononciation.

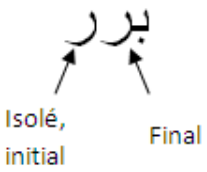
Diacritics Name		Shape	Examples	Diacritics Name		Shape	Examples
Fatha	فتحة	ـَ	مَسْجِدٌ	Sukoon	سكون	ـْ	مَعْرُوفٌ
Thamma	ضمة	ـُ	تَشْكُرُونَ	Shadda	شدة	ـّ	عَنَّا
Kasra	كسرة	ـِ	أَوْلِيَاؤُكُم	Madda	مدة	ـٓ	آبَاؤُكُمْ
Tanween Al-Fath	تنوين الفتح	ـً	قَوْلاً	Alif-Khinjariah	ألف خنجرية	ـٰ	أولئك , إلى إلـه
Tanween Al-Thamm	تنوين الضم	ـٌ	كبيرٌ	Hamzatul-Wasl	همزة وصل	ـٱ	
Tanween Al-Kasr	تنوين الكسر	ـٍ	كأحَدٍ				

TAB. 2.2 – *Liste des diacritiques*
[34]

2.3.2 Les problèmes rencontrés dans la reconnaissance de l'écriture arabe

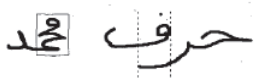
La forme d'écriture de la langue arabe présente quelques problèmes aux développeurs des OCR qui peuvent être résumés dans ce qui suit:

- En texte arabe, les lettres consécutives d'un mot sont connectées suivant la ligne de base utilisée. Pour s'adapter à la ligne de base utilisée, l'alphabet arabe dispose de quatre formes pour chaque lettre (isolée, initiale, médiale, finale). Plusieurs lettres, ز ، و ، ا ، د ، ذ ، ر ne suivent pas cette règle et peuvent avoir différentes formes médiale et finale et quand l'un de ces caractères sera inclus dans le mot alors la lettre la précédant décide de sa forme finale (ou isolée) et lui même décide de sa forme



initiale (ou finale)par exemple dans le mot بر la première lettre a décidé de la forme de la deuxième,et la deuxième a décidé de sa forme et de la forme de la troisième lettre. Cette caractéristique de connectivité cause une difficulté dans l'étape de segmentation. Par rapport aux autres langues, l'écriture arabe est toujours cursive et la vitesse de reconnaissances de ses caractères est faible. [24]

- La plupart des caractères contiennent des points en plus de leurs corps, par exemple la lettre ش qui a le corps de la lettre س avec trois points au dessus, les points peuvent être effacés par erreur dans la phase de prétraitement se qui conduira à des erreurs de reconnaissance. [24]
- En générale le style de l'écriture arabe peut être imprimé(Naskh) ou manuscrit (Ruqq'a), comme on peut trouver d'autre qui sont utilisés pour la calligraphie décorative (Kofi, Thuluth et Diwani). A part les professionnels, d'autres scripteurs ne peuvent pas suivre toutes les règles de l'écriture manuscrite ce qui cause plus de difficulté pour la reconnaissance et rend les bases de données du système volumineuse. [24]
- Quelques caractères dans l'écriture manuscrite peuvent être écrits l'un sur l'autre formant des blocs connectés ou déconnectés créant des chevauchements verticaux ou partager le même espace horizontal se qui augmente la difficulté de segmentation des caractères [25].



2.3.3 Le processus de reconnaissance de l'écriture arabe

La reconnaissance de l'écriture arabe suit les mêmes phases qu'un système OCR que nous présentons dans ce qui suit:

2.3.3.1 Prétraitement

La qualité de reconnaissance dépend de la qualité du texte en entrée et la qualité du texte en entrée dépend de plusieurs facteurs tel que : la résolution du scanner utilisé, le type du document (faxé ou copié) etc. Parmi les opérations de prétraitements généralement utilisées : la binarization, le lissage, le filtrage, la squelettisation, le redressement de l'écriture, la normalisation et la détection de la ligne de base. La figure 2.3 présente quelques opérations de prétraitements réalisés sur l'image du mot محمد

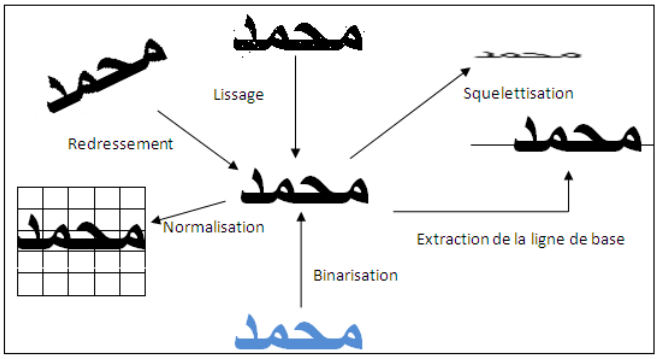


FIG. 2.3 – *Opérations de Prétraitements*

- **La binarisation**: c'est la transformation d'une image en couleur ou en niveaux de gris en une image binaire (composée de deux valeurs 0 et 1) qui permet de distinguer entre le fond d'une image (support papier) et la forme (traits des gravures et des caractères en noir). Les valeurs des niveaux de gris dans l'image qui sont supérieurs à un certain seuil sont mises à 1 et celles inférieurs à ce seuil sont mises à zéro. Le seuil de binarisation doit être choisi attentivement : s'il est très élevé les mots peuvent devenir très mince et peuvent se couper et s'il est très petit les mots peuvent devenir plus épais et des caractères isolés peuvent être connectés [21].
- **Le lissage**: les techniques de lissage permettent de résoudre les problèmes dûs aux artefacts de l'acquisition, tel que les bruits qui conduisent soit à une absence ou à une surcharge de points. Ceci en utilisant les opérations de bouchage et de nettoyage. Les opérations de bouchage égalisent les contours et bouchent les trous internes et celles de nettoyage permettent de supprimer les petites tâches et les excroissances de la forme [30]. Une méthode qui est souvent utilisée pour réduire le bruit dans une image 2D est de parcourir l'image pixel par pixel et changer la valeur du pixel suivant ses 8 voisins (nord, nord-ouest, est,...), un pixel qui est à "un" est mis à "zéro" si il n y' a pas assez de pixels à "un" dans son entourage pour le supporter et vice versa [31]. Plusieurs autres techniques similaires sont utilisées citant les méthodes basées sur la morphologie mathématique [35].
- **Squelettisation** : la squelettisation (amincissement) est une fonction qui permet de réduire l'épaisseur des caractères dans le but de simplifier la représentation des caractères et réduire le temps de traitement. Parmi les algorithmes de squelettisation on trouve dans [36] l'algorithme qui supprime itérativement les points des frontières de la forme jusqu'à obtention d'un squelette d'épaisseur d'un pixel. N.J. Naccache et al [37] présentent une comparaison de performance de 14 algorithmes de squelettisation. L. Lam, S et al [38] présentent un état de l'art sur les méthodologies de squelettisation. Parmi les inconvénients des algorithmes de squelettisation, l'altéra-

tion de la forme des caractères tel que la réduction des points diacritiques doubles et unique à une même forme. Les algorithmes de squelettisation peuvent être classés en deux types : séquentiel [39]et parallèle [40].

- **Extraction de la ligne de base** : la ligne de base contient l'information sur l'orientation du texte et les positions des points de connexion entre les caractères. La méthode la plus utilisée pour trouver la ligne de base est la projection horizontale, qui est un vecteur de taille égale au nombre de lignes de l'image, chaque entrée du vecteur contient le nombre de pixels à "un" dans sa ligne correspondante. La ligne de base apparaît dans l'histogramme comme l'ensemble des entrées consécutives à valeurs maximales. Quelques systèmes AOTR(Arabic Optical Text Recognition) utilisent la ligne de base pour l'alignement des pages penchées, la séparation des lignes dans un bloc de texte et la segmentation des mots en caractères. A. AL-Shatnawi et al [41] présentent une étude et une comparaison des méthodes de détection de la ligne de base et les difficultés rencontrées en traitant l'écriture manuscrite arabe.
- **Normalisation** : la normalisation de la taille des images des caractères consiste à définir dans des matrices de même taille les images de ces derniers, dans le but de faciliter les traitements ultérieurs. Cette opération introduit généralement de légères déformations sur les images, certains traits caractéristiques tels que la ***hampe*** dans les caractères peuvent être éliminés à la suite de la normalisation, ce qui peut entraîner des confusions entre certains caractères [42].
- **Redressement** : l'un des problèmes rencontrés en OCR est l'inclinaison des lignes du texte, qui introduit des difficultés pour la segmentation. L'inclinaison peut être intrinsèque au texte ou provenir de la saisie, si le document a été placé en biais. Il convient alors de le redresser afin de retrouver la structure des lignes horizontales d'une image texte. Si α est l'angle d'inclinaison, pour redresser l'image, une rotation isométrique d'angle -α est opérée grâce à la transformation linéaire suivante tel que X' et Y' représente les nouvelles coordonnés de l'image redressée [42].

$$X' = xcos\alpha - ysin\alpha$$
$$\text{Y'=y cos } \alpha + xsin\alpha$$

2.3.3.2 La segmentation

Cette phase permet de préparer le texte pour être exploiter, en le divisant en unités simples. C'est l'une des phases les plus critiques, difficiles et qui prend plus de temps de traitement. Elle présente un défi dans les systèmes ACR plus que le processus de reconnaissance lui-même.

Les travaux de Parhami et Taraghi [45] en 1981 suivis par ceux d'Amin et Masini [46]

en 1982 sont les premières tentatives pour la segmentation des caractères arabe.

La segmentation peut être implicite ou explicite. Dans la segmentation implicite les caractères sont segmentés lors de la reconnaissance. Le principe de cette technique est d'utiliser une fenêtre mobiles de largeur variable pour fournir des segments provisoires qui serons confirmés ou non par la classification. L'avantage principal de cette technique est qu'elle évite les problèmes recentrés dans la séparation des caractères et n'engendre pas d'erreur de segmentation [43].

A.M. Zeki [44] présente dans son article un état de l'art du problème de segmentation dans la reconnaissance des caractères arabe, dont on tire les méthodes de segmentation explicites suivantes:

1. **Méthodes de segmentation basées sur la projection verticale**: le But de la Projection est de simplifier le système de reconnaissance des caractères par réduction de l'information de deux dimensions à une dimension. Elle s'applique bien aux documents imprimés. Les projections horizontales et verticales sont définies respectivement par les formules:
   $h(i) = \sum(i)p(i,j) \quad et \quad v(j) = \sum(j)p(i,j)$
   Où p(i,j) est la valeur du pixel,elle est égale à " zéro " pour l'arrière plan en blanc et à " un " pour l'écriture en noir , avec i,j désignent respectivement la ligne et la colonne. La projection horizontale est utile pour la séparation des lignes et pour trouver la ligne de base (la plus grande pique de la figure 2.5 représente la ligne de base) alors que la projection verticale aide dans la segmentation du mot en sous mots et caractères. Les figures 2.5 et 2.6 représentent respectivement la projection horizontale et verticale de la phrase de la FiG.2.4

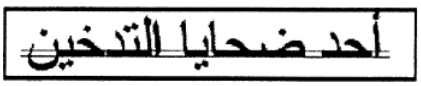


FIG. 2.4 – *Une phrase écrite en arabe*
[44]

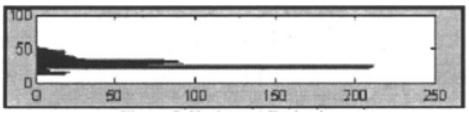


FIG. 2.5 – *La projection horizontale*
[44]

Un algorithme basé sur la projection verticale du mot est développé par Amin (1991) [47] et Altuwaijri et Bagoumi (1994) [48]. Chaque mot est scanné de droite à gauche

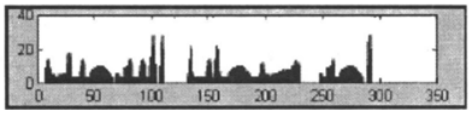


FIG. 2.6 – *La projection vertical*
[44]

et la partie du mot qui a une petite valeur de projection (inférieur à la valeur moyenne de projection) et qui se trouve pré de la ligne de base est marqué comme une bordure potentiel. Quelques règles simples tel que l'épaisseur de la bordure ne doit pas être très petite et l'apparition d' un changement rapide dans la projection vertical aux voisinages de la bordure potentielle sont utilisées pour déterminer les bordures des caractères.

2. **Méthodes de segmentation basées sur la fonction de distance supérieure**: la fonction de distance supérieure est l'ensemble des points les plus élevés dans chaque colonne. Pour chaque fonction de distance supérieure Kurdy and Joukhadar [49] déterminent la ligne de base de chaque sous mot, puis ils mesure la distance entre la ligne de base et le contour supérieur du sous mot et finalement un des trois signes (haut, moyen, bas) est assigné à chaque point. Kurdy et AlSabbagh [50] et Azmi et Kabir [51] ont utilisé la même méthode. Les points voisins ayant les mêmes signes forment un *path*. Si le *path* satisfait quelques conditions tel que la longueur du *path* est supérieur à 1 / 3PS(PS taille du stylo) alors le dernier point du *path* est marqué comme un point de segmentation potentiel.
3. **Méthodes de segmentation basées sur la squelettisation**: dans la reconnaissance de caractères, l'information essentielle sur les formes est sauvegardée dans leurs squelettes. El-Khaly et Sid-Ahmed [52] dans leur méthode, ils trouve d'abord la ligne de base du mot aminci, puis ils ne prennent en considération que les colonnes qui n'ont pas de pixels au dessous ou au dessus de la ligne de base et cela pour trouver les points de segmentation qui seront au milieu du segment de connexion. Jambi, dans sa thèse [24] construit la projection verticale du squelette du mot dont les points diacritiques sont préalablement supprimés. Les points de début et de fin des caractères sont déterminés à partir de la projection verticale ; ces points peuvent être réels ou juste des candidats. Le point de début réel est déterminé si il ya un changement de 0 au non-zéro, alors que le point final réel est déterminé si il ya un changement du non-zéro à 0. Le point de début candidat est déterminé si il ya un changement de 1 à une valeur plus grande, alors que le point final candidat est déterminé si il y a un changement d'une grande valeur à 1. Ce processus sur-segmente le mot et quelques aberrations peuvent être détectées tel que, avoir deux points finals consécutifs, et

c'est difficile de trouver ces points avec certain caractères tel que le caractère س qui a deux points réel et six points candidats de début et finals. Des algorithmes de squelettisation différents peuvent produire différents squelettes de caractères,ceci est l'un des inconvénients des méthodes basées sur la squelettisation des mots, de plus le processus de squelettisation peut altérer la forme du caractère surtout si la qualité de l'image du caractère est médiocre.

4. **Méthodes de segmentation basées sur le tracé du contour (Contour Tracing)**: la segmentation peut être réalisée en traçant le contour externe d'un mot donné. La méthode de segmentation utilisée dans le système SARAT [81] est basée sur le contour externe du corps du mot. Le mot est divisé en série de courbes par la détermination des points de début et de fin du mot,à chaque fois qu'une courbe change de signe d'une valeur +V à une valeur inférieur -V alors le mot sera segmenté à ce point de changement. La méthode d'Al-Ohali [54] se base sur le fait que chaque caractère est formé d'un haut suivi d'un bas ou plat contour. Ainsi, à chaque fois que le contour commence à monter ca indique un point de coupure potentiel. Cette méthode est testée sur des scriptes imprimés, toutefois elle n'été pas parfaite et elle produisait des erreurs. Sari et al [55] ont développé une méthode de segmentation basée sur l'analyse du contour et les règles topologiques. D'abord, ils trouvent les points minimums locaux du contour inférieur de chaque sous mot, puis ils déterminent si les points trouvés sont réellement des points de segmentation en utilisant les règles topologiques. Les minimums locaux du contour supérieur sont détectés dans [56] pour trouver les points de segmentation primaires (PSP) dans lesquels seuls les points de segmentation décisifs (DSP) sont choisis en utilisant les règles suivantes :

   - Le PSP est éliminé s'il est situé au dessus d'une boucle.
   - L'épaisseur de la ligne pour ce PSP doit être plus petite qu'un certain seuil.
   - S'il existe plusieurs DSP dans la même zone de segmentation, le candidat le plus proche de la ligne de base est sélectionné.

   Dans le but de détecter les chevauchements des caractères, deux fonctions sont calculées X(t) et Y(t) pour représenter les variations horizontales et verticales des coordonnées de l'image en fonction du temps. Ces deux fonctions donnent le comportement général de l'écriture sur le contour supérieur. Dans le cas d'accroissement d'amplitude de X(t) en fonction du temps, alors l'écriture est un peu droite et ne présente pas de chevauchement et dans le cas de baisse de l'amplitude du signal X(t) à partir de l'instant k, alors il ya une indication d'un chevauchement.

Les méthodes basées sur le tracé du contour évitent tous les problèmes qui résultent du processus de squelettisation car elles analysent la forme du caractère tel qu'elle est. Wshah dans [57] utilisent une méthode hybride basée sur le squelette et le contour du mot pour la segmentation des mots manuscrits arabe hors ligne.

5. **Méthodes de segmentation basées sur Template Matching**: Bushofa et Spann [58] proposent un algorithme qui cherche l'occurrence d'un angle entre des mots connectés qui se produit à la ligne de base. Premièrement la ligne de base est trouvée, puis l'algorithme procède en balayant l'image de droite à gauche autour de cette ligne avec une fenêtre 7 x 7 pixels et examine le voisinage du pixel central. Le pixel central est candidat à être pris comme point de segmentation. Template matching n'est pas une technique approprié pour la reconnaissance de caractères manuscrits à cause de variations des possibilités de l'écriture des caractères cursifs.
6. **Méthodes de segmentation basées sur les réseaux de neurones** : Hamid et Haraty utilisent dans [25] une méthode hybride de segmentation de texte arabe. Ils combinent entre une méthode heuristique et les réseaux de neurones(RN). Premièrement ils utilisent la méthode heuristique pour scanner le texte manuscrit, extraire les blocs de caractères connectés, générer les caractéristiques topographiques et calculer les points de segmentation qui seront en suite vérifiés et validés par les réseaux de neurones. Pour l'apprentissage du RN, les points de segmentation potentiels seront manuellement classés en points valides ou invalides et sauvegardés dans un fichier avec leurs caractéristiques, ce fichier sera introduit à un RN de type rétropropagation d'erreur, est utilisé pour valider les points de segmentation, ensuite l'erreur entre la sortie désirée et la sortie actuelle peut être déterminée et passée au réseau, une valeur positive indique que le point est un point de segmentation valide et la valeur négative indique que le point est à ignorer.

La segmentation explicite est coûteuse en temps de calcule et engendre des erreurs à cause de la difficulté de trouver les vrais limites des caractères ; ainsi cette technique n'est pas parfaite pour un système de reconnaissance des mots manuscrite [59].

Quelques systèmes de reconnaissance procèdent sans segmentation des mots suivant l'approche holistique où le mot est reconnu entièrement en passant directement à la phase d'extraction de caractéristiques [60–64].

2.3.3.3 Extraction des caractéristiques

C'est l'une des étapes les plus délicates et les plus importantes en OCR. La reconnaissance d'un caractère passe d'abord par l'analyse de sa forme et l'extraction de ses traits caractéristiques (primitives) qui seront exploités pour son identification. L'avantage

principal de cette phase est qu'elle élimine les redondances dans les données. Les caractéristiques peuvent être classées en quatre groupes principaux : caractéristiques structurelles, caractéristiques statistiques, transformations globales et superposition des modèles et corrélation [21].

1. **Caractéristiques structurelles**: les caractéristiques structurelles décrivent une forme en termes de sa topologie et sa géométrie en donnant ses propriétés globales et locales. Elles peuvent tolérer des déformations et des variations dans le style d'écriture et elles décrivent efficacement l'écriture arabe, en particulier manuscrite qui est de nature complexe, cursive, multifontes. Mais leur extraction à partir de l'image n'est pas aussi facile. Parmi ces caractéristiques on peut citer : les boucles, les points d'intersections, les voyellations et les zigzags (hamza), le nombre de points diacritiques et leurs positions par rapport à la ligne de base, la hauteur et la largeur du caractère, la catégorie de la forme (le corps du caractère, point diacritiques et etc.), le nombre de concavité dans les quatre directions principales, la connectivité des caractères etc [65, 66], [26–29].
2. **Les caractéristiques statistiques**: les caractéristiques statistiques décrivent une forme en terme d'un ensemble de mesures extraites à partir de cette forme. Les caractéristiques utilisées pour la reconnaissance de textes arabes sont : le zonage (zonning), les caractéristiques de lieu géométrique (Loci) et les moments. Ali [67] utilise **le zonage** en divisant les caractères en régions chevauchées et non chevauchées et utilise les densités des pixels noirs dans cette région comme des caractéristiques. **La méthode Loci** [68] est basée sur le calcul du nombre de segments blancs et de segments noirs le long d'une ligne verticale traversant la forme, ainsi que leurs longueurs. **La méthode des moments** est l'une des méthodes statistiques les plus utilisées dans le domaine de reconnaissance de forme, les moments d'une forme par rapport à son centre de gravité sont invariants par rapport à la translation et peuvent être normalisé pour être invariants par rapport à la rotation et le changement d'échèle [48], [70, 71]. Cependant quelques chercheurs justifient les limites de leurs systèmes par : le calcule lourd impliqué dans l'approche des moments et leurs sensibilités aux variations dans les entrées [69]. En général, les caractéristiques statistiques peuvent être extraites rapidement et facilement de l'image du texte.
3. **Transformations globales** : la transformation consiste à convertir la représentation de la forme en pixels en une représentation plus abstraite pour réduire la dimension des caractères, tout en conservant le maximum d'informations sur la forme à reconnaître. Une des transformations les plus simples est celle qui représente le squelette ou le contour d'un caractère sous forme d'une chaîne de codes de directions. Les codes de direction peuvent correspondre au huit directions principales

comme dans le code de Freeman. La chaîne de code obtenue est souvent simplifiée pour réduire les redondances et les changements brusques de direction.

Fakir et Sodeyama [72] utilisent la transformation de Hough pour représenter le squelette du caractère comme un ensemble de segments de ligne ensuite utilisent la langueur, l'emplacement et l'inclinaison des segments de ligne comme des caractéristiques. Plusieurs chercheurs utilisent les descripteurs de Fourier dérivés des points de contour des caractères segmentés [73, 74]. Ces descripteurs sont invariants à la translation, rotation et changement d'échèle. Alkhteeb et al [61], appliquent la transformée en cosinus discrète(DCT) sur l'image du mot entier.

En générale, les transformations peuvent être appliquées facilement et tolérer les bruits et variations. Cependant, elles requièrent parfois de les combiner avec des caractéristiques additionnelles pour l'obtention d'un taux de reconnaissance élevé [73]

4. **Superposition des modèles (template matching) et corrélation** : cette technique compare l'image de la forme pixel par pixel à un ensemble de modèles de formes, la forme va appartenir à la classe du modèle à lequel elle ressemble le plus. Cette technique est la plus rapide (lorsque la taille de la base de données n'est pas très grande) mais elle est sensible aux déformations [18], [75, 76].

2.3.3.4 La classification et la reconnaissance

C'est la phase de prise de décision dans les systèmes AOCR. La classification a suivi deux paradigmes principaux : syntactique (structurel) et statistique les réseaux de neurones ont prouvé un troisième paradigme [21].

1. **Approche structurelle**: cette approche repose sur la structure physique du caractère; une forme en entrée est classée en se basant sur ces composantes (primitives) et les relations entre elles. Le classeur identifie les primitives et les analyse suivant un ensemble de règles syntaxiques. La technique de classification la plus populaire dans cette approche est de représenter les caractères avec des règles de production, dans lesquels la partie gauche représente les labels des caractères et la partie droite représente la chaîne de primitives. Les parties droites des règles sont comparées aux chaînes de primitives extraites à partir du mot. Quand il ya une ressemblance la primitive est considérée comme une instance du caractère correspondant [52] [77]. D'autres méthodes représentent un caractère syntaxiquement comme un arbre dont les nœuds internes sont les primitives et les feuilles sont les labels des caractères. Classer un caractère par conséquent revient à trouver un chemin du nœud jusqu'à la feuille [78, 79]. D'autres utilisent la théorie des ensembles flous et modélisent les

caractères manuscrit isolés comme un graphe flou, puis ils comparent le graphe flou des caractères en entrée à ceux des modèles [80]. Les méthodes syntactiques sont spécialement populaires pour la classification du texte manuscrit. L'utilisation de la syntaxe pour exprimer la structure est l'inconvénient de l'approche structurel, puisque les formes peuvent avoir des variations infinies et ne suivent pas toujours les contraintes mathématiques établies par la théorie des langages formels.

2. **Approche statistique**: parmi les méthodes de classification statistiques on trouve celles qui se basent sur les distances, les classeurs bayesiens , les arbres de décision . . . etc. Une forme des classeurs de distance est de comparer les caractéristiques d'une forme avec la valeur moyenne des caractéristiques de chaque classe et étiqueter la forme avec l'étiquette de la classe avec laquelle elle a les plus proches valeurs [81]. Au lieu de comparer à la valeurs moyennes, l'algorithme KNN (K plus proches voisins) affecte une forme inconnue à la classe de son plus proche voisin en la comparant aux formes stockées dans une classe de références nommée prototypes, la forme sera affecté à la classe avec laquelle elle a la plus petite distance [28], [82, 83]. L'un des problèmes de ces méthodes est le choix d'une mesure efficace et précise de distance ou de similarité.

   Une autre méthode statistique est d'utiliser les classeurs bayésiens, qui consistent à choisir parmi un ensemble de caractères, celui pour lequel la suite des primitives extraites ait la plus forte probabilité à posteriori par rapport aux caractères préalablement appris [84]. L'avantage principal des méthodes statistiques est l'apprentissage automatique.

3. **Réseaux de neurones**: est un paradigme émergent pour la reconnaissance de formes. Les primitives extraites sur une image d'un caractère (ou de l'entité choisie) constituent les entrées du réseau.La sortie activée du réseau correspond au caractère reconnu [61], [85], [86], [70], [48], [26], [27]. Le choix de l'architecture du réseau est un compromis entre la complexité des calculs et le taux de reconnaissance.

4. **Modèle Markovien caché (H.M.M)**: c'est une méthode probabiliste qui consiste en un ensemble d'états et les probabilités de transition entre ces états. En plus des observations faites par le système sur une image. Ces dernières sont représentées par des variables aléatoires, dont la distribution dépend de l'état. Elles constituent une représentation séquentielle des caractéristiques de l'image d'entrée [19] [62] [87] [66] [29] .

2.3.3.5 Post-Traitement

Le post-traitement est effectué quand le processus de reconnaissance aboutit à la génération d'une liste de lettres ou de mots possibles, éventuellement classés par ordre décroissant de vraisemblance. Le but principal est d'améliorer le taux de reconnaissance en faisant des corrections orthographiques ou morphologiques à l'aide de dictionnaires. Quand il s'agit de la reconnaissance de phrases entières, on fait intervenir des contraintes de niveaux successifs : lexical, syntaxique ou sémantique. Le post-traitement se charge également de vérifier si la réponse est correcte (même si elle est unique) en se basant sur d'autres informations non disponibles au classeur [21].

2.4 Conclusion

Dans ce chapitre nous avons présenté les systèmes de reconnaissance optique de l'écriture (OCR) et en particulier la reconnaissance de l'écriture arabe(ACR).

Une série d'opérations préalables à la reconnaissance (prétraitement, segmentation, extraction de caractéristiques) sont primordiales à fin de bien préparer les données à reconnaître et d'éliminer les bruits et les redondances. Les ACR dépendent de toutes ces étapes et chaque étape a ses propres effets sur l'efficacité du système, le temps de traitement et de repense et l'erreur de reconnaissance.

Plusieurs travaux traitant l'écriture arabe ont été réalisés, que se soit imprimée ou manuscrite, en ligne ou hors ligne, suivant l'approche analytique ou holistique et en utilisant des techniques de reconnaissance différentes, on trouve les techniques structurelles, statistiques, et celles basées les réseaux de neurones ou les chaînes de Markov.

Ces méthodes ont des avantages mais révèlent des insuffisances envers la nature complexe de l'écriture arabe (cursive, présence de points didactiques, pseudo mots ligaturés,. . .) surtout avec l'écriture manuscrite qui change inter et intra scripteurs ce qui cause un niveaux de difficulté dans la phase de segmentation de l'approche analytique.

Le tableau 2.3 présente les performances et les caractéristiques de quelques systèmes OCR. Nous remarquons que le taux de reconnaissance change d'un classeur à un autre et dépend aussi du type de caractéristiques utilisé. Nous remarquons aussi que le taux de reconnaissance de l'écriture manuscrite hors ligne est faible par rapport à l'imprimé et l'en-ligne.

Pour améliorer le taux de reconnaissance et remédier aux problèmes rencontrés, la tendance est de construire des systèmes hybrides qui utilisent plusieurs types de caractéristiques et combinent entre les classifieurs et d'utiliser et tirer partie des avantages

des algorithmes bio-inspirés (systèmes immunitaires, algorithmes génétique, les colonies de Fourmis, ...) qui ont l'avantage de simuler les systèmes biologiques réels, qui seront l'objet du chapitre prochain.

Référence	**Le classeur**	**Type d'écriture**	**Caractéristiques**	**Segmentation**	**Performances**
M.Sarfraz et al 2003 [70]	Réseaux de neurones RBF	Texte imprimé, hors ligne	Statistique (moments de HU)	Oui	73%
M.Altuwaijri et al 1994 [48]	Réseaux de neurones PMC	Texte imprimé, hors ligne	Statistique (moments de HU)	Oui	90 %
A.Broumandnia et al 2007 [88]	Réseaux de neurones	Texte manuscrit hors ligne	Transformation globale (transformée en ondelette)	Non	95,8%
R. Haraty et al 2004 [26]	Réseaux de neurones PMC	Texte caractère manuscrit hors ligne	Structurel (points finale/branchement/croisement, boucle, longueur, largeur)	Oui	73%
L. Souici-Meslati et al 2004 [27]	Réseaux de neurones +Approche symbolique	Montants littéraux	Structurel (boucle, nombre de ascendeurs, nombre de descendeurs, point diacritiques, composantes connectés)	Non	93%
J.H . AlKhateeb 2008 [61]	Réseaux de neurones	Caractères manuscrits hors ligne	Transformation globale (transformé en DCT)	Non	82.5%
M. Kherallah et al [89]	Algorithmes génétique	Mots manuscrits en ligne	Chaîne de code	Non	97%
M.Pechwitz et al 2003	Chaîne de Markov	Mots manuscrits hors ligne	Statistique (Densité des pixels)	NON	89%

Référence	**Le classeur**	**Type d'écriture**	**Caractéristiques**	**Segmentation**	**Performances**
M.S. Khorsheed 2003 [66]	Chaîne de Markov	Caractères manuscrits hors ligne	Structurel (Boucles, points diacritiques, points finale/branchement/croisement)	Non	87%
R. Safabakhsh et al 2005 [29]	Chaîne de Markov	Caractère manuscrit hors ligne	Structurel et statistique (Descripteur de Fourier, le nombre de boucles, le rapport longueur /largeur, la position des connections droite et gauche et la densité des pixels)	Oui	86 ,17%
M.Fakir et al 2009 [72]	Programmation dynamique template matching	Texte imprimé hors ligne	Transformation globale (Transformé de haugh)	Oui	95%
A.Amin 2000 [28]	Arbre de décision	Texte imprimé hors ligne	Structurel (Caractéristiques globales (nombre de sous mots, nombre de boucles, nombre de peak , l'emplacement des caractères supplémentaires)	Non	92%
Abdelazim et al 1990 [68]	Classeur Bayesien	Texte imprimé hors ligne	Statistique (Caractéristique Loci)	Oui	98%
A.Amine et al 1980 [28]	KNN	Texte imprimé en ligne	Chaîne de code	Non	95.4%

TAB. 2.3 – ***Caractéristiques et performances de quelques systèmes OCR***

Chapitre 3

Les algorithmes bio-inspirés

Pendant les derrières décennies il y a eu un grand intérêt d'introduire la biologie comme source d'inspiration pour le développent de solutions à différents problèmes. D'où la naissance des algorithmes bio-inspirés tel que les systèmes immunitaires artificiels, les algorithmes génétiques, réseaux de neurones, les fourmis artificiels ... etc.

3.1 Les systèmes immunitaires artificiels (AIS)

Parmi les algorithmes inspirés de la biologie on trouve les systèmes immunitaires artificiels (AIS) dont les principes sont issues des systèmes immunitaires naturels.

3.1.1 Systèmes immunitaires naturels

Dans cette partie nous présenterons quelques généralités sur les systèmes immunitaires naturels

3.1.1.1 Définition

Un système immunitaire est un système de défense, il protège le corps des vertébrés contre des agents infectieux tel que les virus, bactéries et d'autres parasites.

Il existe deux types d'immunités inter-reliés :

- **Immunité innée :** elle est appelée ainsi car le corps naît avec la capacité de reconnaître certain microbes et les détruire immédiatement et elle est non spécifique. Elle joue un rôle principale dans l'initiation et la régulation de la reponse immunitaire mais elle n'est pas une solution complète de protection du corps [90].

- **Immunité acquise :** elle permet au système immunitaire de lancer une attaque sur des intrus que l'immunité innée ne reconnaît pas, elle est caractérisée par sa spécificité.Dans le processus de reconnaissance et de destruction des antigènes, cette immunité se sert de deux types de lymphocytes distribués(les cellules B et T) et se base sur la sélection clonale. Les anticorps jouent le rôle principal dans ce type de système. Les substances capable de déclencher une réponse des lymphocytes sont appelés antigènes [91].

La réponse immunitaire peut être classée en réponse primaire et secondaire : [92]

- **Réponse primaire :** elle est provoquée lorsque le système immunitaire rencontre l'antigène pour la première fois, un nombre d'anticorps sera produit en réponse à l'infection qui aide à l'élimination de l'antigène, le nombre d'anticorps va diminuer au courts du temps jusqu'à ce qu'il rencontre l'antigène une autre fois.
- **Réponse secondaire :** elle est spécifique à l'antigène qui a déjà initié la réponse immunitaire. Elle cause une croissance très rapide de la quantité de cellules B et des anticorps. Cette réponse est rapide et dû aux cellules mémoires qui se souvient de l'antigène, ainsi l'immunité n'a pas besoin d'être élaborée car elle existe déjà.

3.1.1.2 Les cellules immunitaires

Les cellules qui participent aux réponses immunitaires sont nombreuses. Les plus importantes sont les cellules lymphoïdes (lymphocytes B, T, NK Natural Killer), les macrophages, les cellules mononuclées, les cellules dendritiques et les granulocytes. Seules les lymphocytes T et B possèdent les caractéristiques de la réponse immunitaire adaptative (acquise, spécifique). Les autres cellules jouent des rôles accessoires (réponse immunitaire innée): activation des lymphocytes, augmentation de l'élimination de l'antigène, sécrétion d'effecteurs du système immunitaire.

- **Les lymphocytes B et les anticorps :** la fonction principale des lymphocytes B est la production des anticorps comme réponse aux protéines exogènes, chaque cellule B est programmée à produire des anticorps spécifiques qui reconnaissent et attachent un autre protéine particulier. La production des anticorps est souvent la manière de signaler les cellules à tuer, ingérer ou enlever la substance attachée [93].
- **Les lymphocytes T et les lymphokines :** pour que les lymphocytes T reconnaissent un antigène, il faut qu'il soit préparé par d'autres cellules, essentiellement les CPA, puis présenté à leur surface sous la forme d'un complexe peptide-CMH, puis elles sécrètent des substances appelées lymphokines qui constituent des messagers chimiques puissants qui aident les cellules B dans la stimulation des antigènes [93].

- **Phagocytes:** les phagocytes sont des cellules capables d'ingérer et de digérer les microorganismes et les particules antigéniques. Quelques phagocytes ont la capacité aussi de présenter les antigènes aux lymphocytes, ainsi sont appelées cellules de présentation d'antigènes(APC). Les phagocytes les plus importantes sont les macrophages et les monocytes.Les monocytes circulent dans le sens et migrent aux tissus, où elles deviennent des macrophages (grand mangeur), elles présentent l'antigène aux lymphocytes, elles jouent un rôle important au début de la réponse immunitaire [93].

3.1.1.3 Caractéristiques du système immunitaire

Le systéme immunitaire a de nombreuses caractéristiques qui peuvent âtre résumés comme suit [99, 100] :

- **Reconnaissance :** le système immunitaire a la capacité de reconnaître, identifier et répondre à un vaste nombre d'antigènes différents. En plus, il peut faire la différence entre les cellules de l'individu et les cellules étrangères.
- **Extraction de caractéristiques :** par l'utilisation de cellules de présentation d'antigène (APC), le système immunitaire a la capacité d'extraire les caractéristiques de l'antigène, avant d'être présentée à d'autres cellules immunisées, y compris les lymphocytes.
- **Diversité :** il y a deux processus principaux impliqués dans la génération et le maintien de la diversité dans le système immunitaire. Le premier est la génération des molécules de récepteur par la recombinaison des segments de gènes à partir des bibliothèques de gènes. En recombinant des gènes d'un ensemble fini, le système immunitaire est capable de produire un nombre presque infini de types de récepteurs, de ce fait le système immunitaire est doté d'une grande assurance dans l'univers des antigènes. Le deuxième processus, qui assiste la diversité dans le système immunitaire est la hyper mutation somatique. Les cellules immunisées se reproduisent en réponse aux antigènes envahissants le corps. Pendant cette reproduction, elles sont soumises à un processus de mutation somatique avec des taux élevés qui permettent la création de nouveaux modèles de molécules de récepteurs, ainsi la diversité des récepteurs immunisés augmente [94].
- **Apprentissage :** le mécanisme de l'hyper mutation somatique suivi d'un mécanisme de sélection permet également au système immunitaire d'améliorer sa réponse à un pathogène envahissant le corps. Ce processus est nommé la maturation d'affinité [95]. Cette dernière garantit que le système immunitaire améliore sa tâche de reconnaissance des formes d'antigène.

- **Mémoire :** après une réponse immunitaire à un antigène donné, quelques ensembles de cellules et molécules seront dotés d'une grande durée de vie afin de fournir une réponse immunitaire plus rapide et plus puissante à des futures infections par le même antigène ou des antigènes semblables. Ce processus, connu sous le nom de maturation de la réponse immunitaire, elle permet le maintien de ces cellules et molécules lors de la reconnaissance de l'antigène. C'est le principe fondamentale des procédures de vaccination dans la médecine et l'immunothérapie. Un échantillon affaibli ou mort d'un antigène (par exemple, un virus) est inoculé dans un individu afin de lancer une réponse immunitaire (sans les symptômes de la maladie) afin de produire des cellules et des molécules de mémoire à cet antigène.
- **Détection distribuée :** il y a une distribution inhérente dans le système immunitaire. Le contrôle n'est pas centralisée, chaque cellule immunisée est spécifiquement stimulée et répond aux nouveaux antigènes dans n'importe quel endroit.
- **Autorégulation :** la dynamique du système immunitaire est telle que la population de système immunitaire n'est pas contrôlée par des interactions locales mais pas par un point central de contrôle. Après qu'une maladie ait été combattue avec succès, le système immunitaire revient à son état d'équilibre normal, jusqu'à ce qu'il soit nécessaire de répondre à un autre antigène. La théorie du réseau immunitaire explique explicitement ce type de mécanisme autorégulateur.
- **Métadynamique :** le système immunitaire crée constamment de nouvelles cellules et molécules et élimine celles qui sont trop vielles ou qui ne sont pas d'une grande utilité [96].

3.1.1.4 Fonctionnement du système immunitaire

Quand un microbe pathogène étranger infectieux attaque le corps, le système immunitaire suit les étapes principales de défense présentés dans la figure 3.1 [93] qui sont:

1. Les cellules de représentation d'antigène (APC) spécialisées comme les macrophages ingèrent et digèrent les antigènes qu'elles rencontrent et les fragmentent en peptides antigéniques [98].
2. Les morceaux de ces peptides sont attachés au complexe majeur d'histocompatibilité (CMH) et sont exposés sur la surface de la cellule. Les lymphocytes T ont des récepteurs qui leurs permettent la reconnaissance des différentes combinaisons de peptides-CMH.
3. Les cellules T activées sécrètent les lymphokines, (ou des signaux chimiques) qui mobilisent les autres composantes du système immunitaire.

4. Les lymphocytes B qui ont aussi des récepteurs d'une spécificité unique sur leurs surfaces répondent à ces signaux contrairement aux récepteurs des cellules T, ceux des cellules B peuvent reconnaître les parties de l'antigène sans les molécules CMH.
5. Quand les cellules B deviennent activées, elle se différencient en cellules plasma qui sécrètent les protéines d'anticorps qui sont des formes solubles de leurs récepteurs.
6. En se liant à l'antigène trouvé, les anticorps peuvent les neutraliser ou précipiter leur destruction avec des enzymes ou par des cellules de balayage. Quelques cellules T et B deviennent des cellules mémoire ce qui augmente la rapidité du système immunitaire dans l'élimination du même antigène s'il se présente encore dans le future.

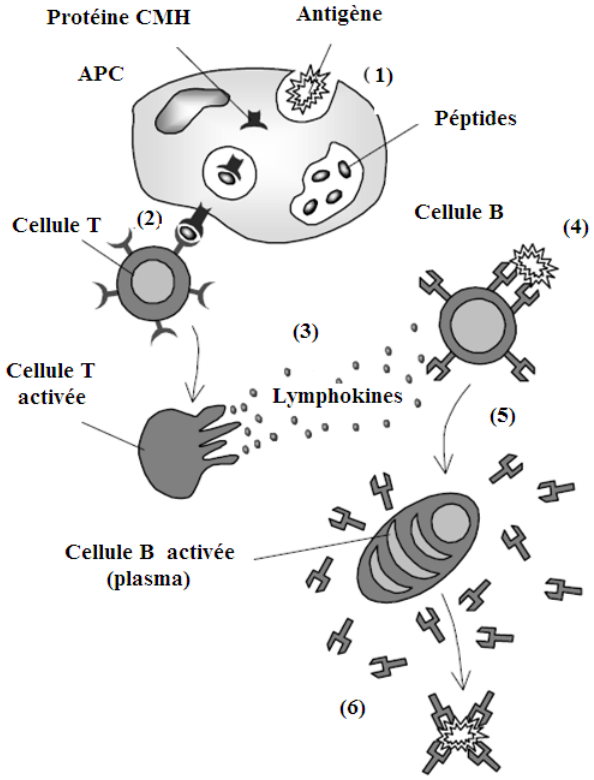


FIG. 3.1 – *Fonctionnement du système immunitaire*
[93]

3.1.2 Caractéristiques des AIS

Ils sont des systèmes adaptatifs, inspirés de l'immunologie théorique, des fonctions, des principes et des modèles immunitaires qui sont appliqués à la résolution des problèmes [100]. Leurs origines reviennent aux travaux théoriques d'immunologie [96, 97].

Les AISs visent à développer un modèle informatique qui préserve les caractéristiques

importantes des SINs telles que : la reconnaissance, la discrimination, la mémorisation, l'apprentissage, l'auto-organisation, l'adaptation, la distribution, la robustesse et l'évolutivité.

3.1.3 Composantes principales des AIS

En générale, l'ingénierie immunitaire a trois composantes principales: un modèle pour la représentation de la problématique, un mécanisme pour l'évaluation de l'affinité entre les anticorps et les antigènes et un algorithme pour le contrôle du système (sélection clonale, sélection négative, réseaux immunitaire) [101].

3.1.3.1 Modèles de représentation de données

Pour chaque problème à résoudre, un modèle de représentation de ses données doit être établit. Définir les parties des données qui seront représentés par les anticorps et celles qui seront représentés par les antigènes, ainsi le codage à leurs appliquer (réel, entier, binaire , caractères , symbolique ..) qui est très important pour le succès de l'algorithme à utiliser.

Les anticorps et les antigènes sont définis comme suit [102]:

- **Anticorps:** c'est un vecteur de caractéristiques couplé avec la sortie ou la classe qui lui est associée, la combinaison en sortie du vecteur de caractéristiques est référé à un anticorps s'il appartient à un ARB.
- **Antigène :** il a la même représentation que celle des anticorps, cependant, la combinaison de la calasse en sortie du vecteur de caractéristique est référé à un antigène lorsqu'il est présenté au ARBs pour stimulation et réponse. C'est une solution à un problème donnée.
- **Vecteur de caractéristiques** : c'est une instance de donnée représentée comme une séquence de valeurs. Chaque position dans la séquence représente une caractéristique différente associée à une donnée, et chaque caractéristique a son propre intervalle de valeurs.
- **ARB :** il se compose d'anticorps, du nombre de ressources tenus par une cellule immunitaire et la valeur courante de stimulation de la cellule.

3.1.3.2 Un mécanisme de calcule d'affinité

L'affinité est le moyen de quantification de l'interaction entre les éléments du système. Elle mesure la similarité entre deux anticorps et/ou antigènes. Plus l'affinité est faible plus le degré de similarité est élevé. La méthode de calcule de l'affinité dépend du

type d'encodage de données utilisé. Timis et al [101] proposent deux méthodes différentes pour calculer l'affinité. La première, nommée recognition ball, elle consiste à calculer la distance entre chaque antigène et les récepteurs(anticorps) qui les ont détectés. Si cette distance est comprise dans le rayon "'ϵ "'alors l'affinité est considérée comme suffisante pour générer une réponse immunitaire. Cette distance peut être obtenue en utilisant les distances Euclidienne, de Manhattan ou de Hamming.

Soit Ab= (Ab1, Ab2, ..., AbL) les coordonnées de l'anticorps et Ag = (Ag1, Ag2, ..., AgL) celles de l'antigène.

Recognition ball : $B_\epsilon(Ab) = \{Ag | D(Ab,Ag) < \epsilon\}$

- **La distance Euclidienne :** $D(Ab,Ag) = \sqrt{\sum_{i=1}^{L}(Ab_i - Ag_i)^2}$
- **La distance de Hamming:** $D(Ab,Ag) = \sum_{i=1}^{L} |Ab_i - Ag_i|$
- **La distance de Manhattan:** $D(Ab,Ag) = \sum_{i=1}^{L} \delta$ où $\delta = \begin{cases} 1 si Ab_i = Ag_i \\ 0 sinon \end{cases}$

La seconde méthode exposée dans [101] est la correspondance r-contiguous c'est-à-dire que "deux éléments ont la même longueur de correspondance si au moins r caractères contigus sont identiques". Dans l'exemple de la figure 3.2, il y a deux vecteurs de huit paramètres dont leur valeur est soit 0 ou 1. Le calcul de l'affinité se fait en appliquant un ou exclusif entre les deux vecteurs. L'affinité entre ces deux vecteurs est de 4 car il y a quatre valeurs contiguës à 1.

0	0	1	1	0	0	1	1
0	1	1	0	1	1	0	1
1	1	0	1	1	1	1	0

FIG. 3.2 – *Affinité en nombre de bits contigus*

3.1.3.3 Un mécanisme de contrôle du système

Pour la résolution d'un problème donné, les AIS offrent des algorithmes différents tel que la sélection clonale, les sélections négative et positive, les réseaux immunitaires et la théorie du danger qui sont conçus pour différentes applications, le choix de l'algorithme à utiliser dépend du problème à résoudre. Ces algorithmes seront détaillés dans la section suivante.

3.1.4 Les algorithmes existant et leurs applications

Les algorithmes des AIS peuvent être divisés en trois grandes parties, chacune d'elles s'inspire d'un comportement ou d'une théorie du système immunitaire biologique : la sélection clonale, la sélection négative et positive, les réseaux immunitaires et la théorie

du danger. Emma.h a présenté dans [103] un état de l'art sur les domaines d'application des AIS qui sont très variés et ceci grâce à la polyvalence et la puissance du système immunitaire biologique duquel ils s'inspirent.

3.1.4.1 La sélection clonale

Cet algorithme utilise la propriété de la sélection clonale réalisée par les cellules immunitaire,ainsi que la maturation d'affinité [104]. Le principe générale de cet algorithme est présenté dans l'algorithme présenté dans la figure 3.3.

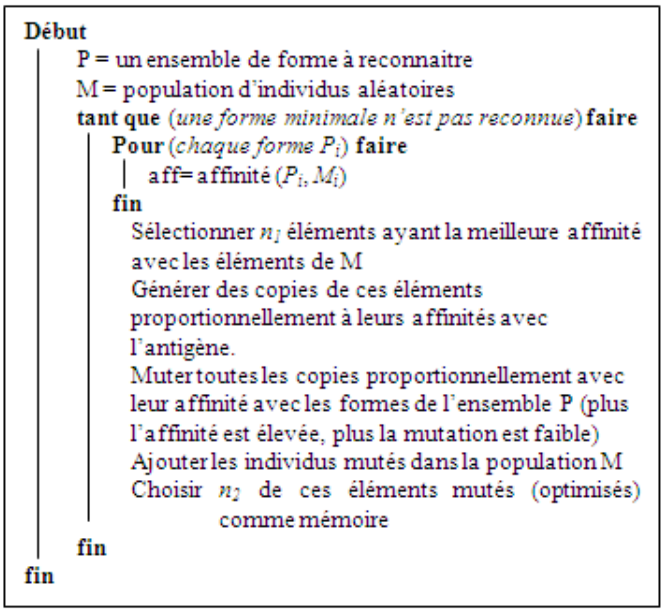


FIG. 3.3 – *Algorithme de la sélection clonale*

Les algorithmes de sélection clonale sont le plus souvent utilisés : dans des applications d'optimisation vu que les cellules B deviennent de plus en plus affines aux antigènes, dans des applications de détection d'intrusions également où l'on ne peut répertorier tous les éléments indésirables et où ces éléments sont extrêmement variés, ainsi que d'autres applications comme la reconnaissance de caractères.

3.1.4.2 La sélection négative

Cet algorithme est inspiré de la sélection négative immunitaire. Il se base sur des détecteurs qui peuvent reconnaître n'importe quelle forme de l'ensemble du soi et détecter les données anormales [105]. IL est présenté dans l'algorithme présenté dans la figure 3.4.

```
début
    S = Un ensemble d'éléments du soi
    D = Un ensemble de détecteurs
    SeuilAff = seuil d'affinité
    tant que (i < nbDetecteurs) faire
        Générer un détecteur d_i de manière à ce qu'il n'ait pas d'affinité avec un
        élément de S
        si (affinite(d_i, S_i) > SeuilAff) alors
            classer S_i comme non-soi
        sinon
            classer S_i comme soi
        fin
    fin
    retourner D Un ensemble de détecteurs
fin
```

FIG. 3.4 – *Algorithme de la sélection négative*

Les algorithmes de sélection négative sont le plus souvent utilisés dans des applications de détection d'intrusion at de fraude, ainsi que des détecteurs d'anomalies.

3.1.4.3 Les réseaux immunitaires

Une autre classe d'algorithmes immunitaires est les réseaux immunitaires. N.Jerne [106], en 1974 a supposé que les cellules et les molécules immunitaires sont capables de s'identifier en plus de reconnaître les antigènes envahissant. Ces cellules stimulent et se suppriment d'une manière à mener à la stabilité du réseau. Deux cellules se connectent si l'affinité qu'elles partagent dépasse un certain seuil. La puissance de connexion entre elles est proportionnelle à l'affinité qu'elles partagent. Le principe générale de cet algorithme est présenté dans l'algorithme présenté dans la figure 3.5.

1. **Initialisation :** initialisation du réseau de cellules immunitaire.
2. **Pour** chaque antigène **faire** :
   - 2.1 **Reconnaissance antigénique :** apparier les cellules du réseau avec l'antigène.
   - 2.2 **Interaction du réseau :** apparier les cellules du réseau entre elles.
   - 2.3 **Métadynamique :** introduire de nouvelles cellules au réseau et éliminer celles qui sont inutiles (suivant un certain critère).
   - 2.4 **Niveau de stimulation :** évaluer le niveau de stimulation de chaque cellule du réseau en prenant en comptes les résultats des étapes précédentes.
   - 2.5 **Dynamique du réseau :** mettre à jour la structure du réseau suivant le niveau de stimulation des cellules.
3. **Répéter :** l'étape 2 jusqu'à convergence au critère d'arrêt.

FIG. 3.5 – *Algorithme du réseaux immunitaire*

Les réseaux immunitaires sont beaucoup plus appliqués dans le domaine de la reconnaissance de forme, la classification et la robotique.

3.1.5 Conclusion

Les systèmes immunitaires artificiels sont des systèmes de résolutions puissant, leurs force et puissance reviennent aux caractéristiques qu'ils engendrent tel que la reconnaissance, la mémoire, la distribution , l'auto organisation et etc. Ils ont plusieurs techniques et algorithmes, chacun mappe un mécanisme différent du système immunitaire naturel(sélection clonale, sélection négatives, ...), et utilisé pour la résolution de différents types de problèmes. Ils sont très intéressants pour la résolution de problèmes dans des environnements dynamiques et distribués ainsi que les problèmes de reconnaissance de formes, d'optimisation, de détection d'anomalies et des problèmes de datamining tel que la classification, le clustering et etc. Malgré le succès des techniques des AIS il reste encore des issus ouverts et il n'y a pas encore une architecture unifié qui intègre les différents modèles des AIS.

3.2 Les réseaux de neurones

Parmi d'autres algorithmes s'inspirant de la biologie , on trouve les réseaux de neurones artificiels(RN) qui représente une tentative de reproduire artificiellement le fonctionne-

ment du cerveau humain. En 1943, Warren McCulloch et Walter Pitts, en s'inspirant de leurs travaux sur le neurone biologique, ont proposé un des premiers modèles de neurone artificiel [107] qui deviendra la base des réseaux de neurones artificiels.

3.2.1 Le neurone biologique

Le cerveau humain, est le meilleur modèle de la machine polyvalente rapide et surtout douée d'une énorme capacité d'auto organisation. Il est constitué d'un grand nombre de cellules nerveuses appelées "neurones" [108]. Ces dernières sont constituées de trois parties essentielles : le corps cellulaire, les dendrites et l'axone. La figure 3.6 représente un schéma générale d'un neurone biologique.

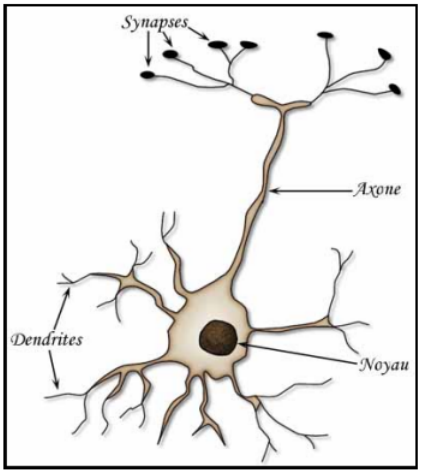


FIG. 3.6 – *Structure d'un neurone biologique*

3.2.2 Le neurone formel

Le neurone formel est le modèle mathématique du neurone biologique. Il fait la somme pondérée de ses entrées, suivie d'une non linéarité(élément de décision pour les classifieurs) appelée fonction d'activation ou fonction de seuil. Les entrées d'un neurone sont soit des entrées externes, soit des sorties d'autre neurones. La figure 3.7 présente le schéma générale d'un neurone artificiel.

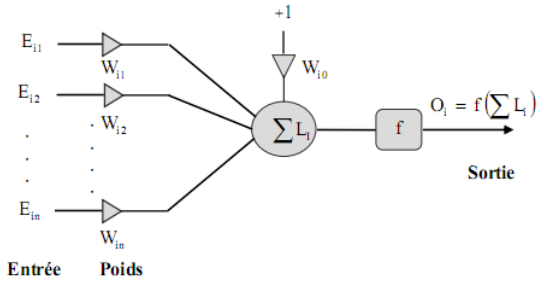


FIG. 3.7 – *Structure d'un neurone formel*

L'équation de sortie O_i du neurone i est donné par :

$O_i = f(L_i)$.

où $L_i = \sum W_{ij}^k - W_{i0}^k$

Les coefficients de pondération W_{ij} sont appelés coefficients synaptiques.

Souvent, il y a un paramètre additionnel W_{i0}, ce terme est considéré comme la valeur du seuil interne du neurone.

- **Fonction d'activation**

  C'est une fonction permet de définir l'état interne du neurone en fonction de son entrée totale. Les fonctions les plus souvent utilisées sont représentées par la figure 3.8.

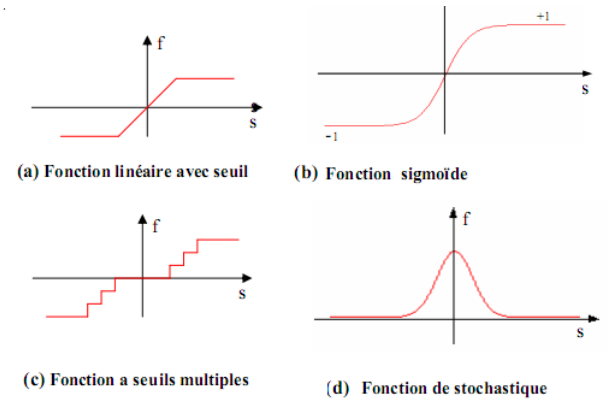


FIG. 3.8 – *Les différentes formes de la fonction d'activation*

3.2.3 Propriétés des réseaux de neurones

Un réseaux de neurones est un réseau composé de plusieurs neurones formels. Ces derniers sont interconnectés entre eux, de sorte que les signaux sortant(outputs) des neurones deviennent des signaux entrant (inputs) d'autres neurones.

L 'intérêt porté aujourd'hui aux réseaux de neurones tient sa justification dans quelques propriétés intéressantes qu'ils possèdent et qui devraient permettre de dépasser les limitations de l'informatique traditionnelle, tant au niveau programmation qu' au niveau machine [109] parmi ces propriétés :

- **Le parallélisme**: qui implique un traitement rapide et une tolérance au failles matériels.
- **La capacité d'adaptation et d'apprentissage** : permet au système de mettre à jour (modifier) sa structure interne pour répondre aux changements dans l'environnement.
- **La mémoire distribué**: dans les réseaux de neurones, la mémoire correspond à une carte d'activation de réseaux. Cette carte est en quelque sorte un codage d'un fait mémorisé ce qui attribue à ce réseaux l'avantage de résister aux bruits (pannes) car la perte d'un élément ne correspond pas à la perte d'un fait mémorisé.
- **La capacité de généralisation** : permet l'application du modèle aux données non apprises.

3.2.4 Modèles des réseaux de neurones

Un réseau de neurones est caractérisé par sa topologie qui dépend de la façon dont les neurones sont reliés (réseaux en couche, complètement connecté et récurrent), par sa fonction d'activation et par le mode d'apprentissage utilisé (supervisé, non supervisé). Il existe plusieurs modèles de RN reflétant les différentes topologies, nous citons ici quelques modèles.

3.2.4.1 Modèle de Kohonen

Le modèle de kohonen appelé aussi réseaux auto-organisés(SOM: self-organizing maps) [110], est une variante très importante des réseaux de neurones pour la segmentation et la compression de données plutôt que pour la classification. Les réseaux de kohonen ont une architecture à deux couches, une couche d'entrée et une couche de sortie, chaque neurone de la couche d'entrée est complètement connecté aux neurones de la couche de sortie. L'apprentissage de ce type de réseaux est non-supervisé et compétitif. Les poids des liaisons entre neurones sont mis à jour à chaque arrivée d'un individu de l'ensemble d'apprentissage.

3.2.4.2 Modèle de Hopfield

C'est un réseau monocouche complètement connecté, chaque neurone est connecté aux autres neurones du réseau et il n'y a aucune différence entre les neurones d'entrée et de sortie. La matrice des poids des connexions est symétrique et nulle sur la diagonale(pas de connexions bouclées) [111]. Les réseaux de Hopfield se comportent comme des mémoires associatives non-linéaire, ils sont capable de restituer une information stockée et apprise par le réseau à partir des données bruitées ou incomplètes. Il sont appropriés à la reconnaissance de caractères et appliqués aussi à la résolution des problèmes d'optimisation.

3.2.4.3 Le perceptron multicouches

Le perceptron multicouches [112](MLP multilayer perceptron) est un réseau orienté de neurones artificiels organisé en couches où l'information circule dans un seul sens, de la couche d'entrée vers la couche de sortie. Un perceptron multicouches contient généralement trois couches, une couche d'entrée, une couche de sortie et une couche intermédiaire appelé couche cachée. Les neurones de la couche d'entrée sont passifs, ils n'altèrent pas les informations, ils reçoivent les valeurs des attributs et les transmettent à la couche cachée. Cette dernière est la couche la plus importante dans le MLP, ses neurones sont connectés aux neurones des autres couches, c'est la couche responsable de l'apprentissage du réseau. Ce modèle de réseau de neurones est très utilisé pour la classification de données et l'architecture du réseau dépend de la nature du problème, le nombre d'attributs de classification détermine le nombre de neurones de la couche d'entrée et la couche de sortie aura autant de neurones que de classes à attribuer au individus.

3.2.5 Conclusion

Les réseaux de neurones artificiels sont des outils puissants capables d'être utilisés dans plusieurs domaines tel que le traitement du signal, l'aide à la décision, la robotique, la reconnaissance des formes ...etc. Ils ont des propriétés qui les ont rendu des outils standard dans le data mining. Plusieurs modèles de RN ont été présentés chacun a ses particularités et ses champs d'application, cependant ils présentent quelques limitations à savoir leur performance dépend de la qualité et la quantité des données traitées, ils ont un faible pouvoir explicatif (boite noire), manque de règles claires ou de directives fixes pour la conception d'une architecture des RN optimales.

3.3 Les algorithmes génétiques

Les algorithmes génétiques (AG) sont une autre famille des algorithmes inspirés de la biologie. Ils sont des algorithmes d'exploration fondés sur les mécanismes de la sélection naturelle et de la génétique. Ils se basent sur la théorie de Darwin " survie des individus les mieux adaptés ". Ces algorithmes sont initialement introduits par John Holland [113] qui a développé leurs principes fondamentaux, puis Goldberg [114] les a utilisés pour résoudre des problèmes d'optimisation.

Les AG ont prouvé leur succès dans les problèmes d'optimisation à large espace de solutions [115]. Ils sont utilisés lorsque la recherche exhaustive d'une solution est coûteuse en termes de temps d'exécution.

3.3.1 Terminologies

Les algorithmes génétiques utilisent une terminologie qu'ils inspirent de la génétique ainsi que d'autres notions qui sont propres à leur domaine.

La terminologie empruntée de la génétique :

- Les chromosomes sont les éléments à partir desquels sont élaborés les solutions (individus).
- La population est l'ensemble des chromosomes.
- La reproduction est l'étape de combinaison des chromosomes, la mutation et le croisement génétiques sont des méthodes de reproduction.

Les notions propres au domaine des AG :

- L'indice de qualité (fitness), aussi appelé indice de performance, est une mesure abstraite permettant de classer les chromosomes.
- La fonction d'évaluation ou fonction coût est la formule théorique qui permet de calculer l'indice de qualité d'un chromosome.

3.3.2 Caractéristiques:

Les AG sont caractérisés par :

- Ils utilisent un codage des paramètres, et non les paramètres eux-mêmes.
- Ils travaillent sur une population d'individus, au lieu d'un individu unique.
- Ils utilisent une fonction d'évaluation.
- Ils utilisent des règles de transition probabilistes et non déterministes.

3.3.3 Principe de fonctionnement

Les AG sont fondés sur [116]:

- Une représentation chromosomique des solutions du problème.
- Une méthode pour générer une population initiale de solutions.
- Une méthode d'évaluation qui classe les solutions selon leurs aptitudes.
- Des opérateurs génétiques, qui définissent la manière dont les caractéristiques des parents sont transmises aux descendants.

3.3.3.1 Représentation des données

Goldberg [114] donne deux principes de base pour choisir la représentation des solutions d'un AG. Tout d'abord, le codage doit contenir des blocs de construction contenant une information ayant du sens. Ensuite, l'alphabet doit être le plus petit permettant une expression naturelle du problème. Le choix du codage des données dépend de la spécificité du problème traité et il conditionne fortement l'efficacité de l'AG.On distingue trois type de codage: numérique, symbolique, alphanumérique.

3.3.3.2 Génération de la population initiale

Le choix de la population initiale d'individus influence fortement la rapidité de convergence de l'AG.Si la position de l'optimum dans l'espace des solutions est totalement inconnue, il est naturel de générer aléatoirement des individus en faisant des tirages uniformes dans chacun des domaines associés aux composantes de l'ensemble de solutions. Si des informations a priori sur le problème sont disponibles, il paraît bien évidemment naturel de générer les individus dans un sous domaine particulier afin d'accélérer la convergence.

3.3.3.3 Fonction d'évaluation

La fonction d'évaluation quantifie la qualité de chaque chromosome(solution) par rapport au problème. Elle est utilisée à fin de sélectionner les chromosomes pour la reproduction. Les chromosomes ayant une bonne qualité auront plus de chance d'être sélectionnés pour la reproduction et donc plus de chance que la population suivante hérite de leur matériel génétique. La fonction d'évaluation produit la pression qui permet de faire évoluer la population de l'AG vers des individus de meilleur qualité. Le choix de cette fonction va fortement influer sur le succès de l'algorithme génétique.

3.3.3.4 Les opérateurs

Les opérateurs utilisés par les AG sont la base des AGs. Ils définissent la manière dont les individus se recombinent et s'agencent pendant la phase de reproduction. On distingue différents opérateurs: sélection, croisement et mutation.

1. **Sélection**

   Cette opération est fondée sur le principe d'adaptation de chaque individu d'une population à son environnement, suivant la théorie de la sélection naturelle introduite par Charles Darwin. Ainsi, seuls les individus ayant des coûts de performance élevés seront sélectionnés à survivre et à se multiplier.

   Il existe plusieurs méthodes de sélection. Les plus connues sont la sélection proportionnelle à la fonction fitnesse , la sélection par la roulette , la sélection sur le rang et la sélection en tournoi.

   - **Sélection par la roulette (*wheel*)**: elle s'inspire des roues de loterie. A chaque individus de la population est associé un secteur d'une roue. L'angle du secteur étant proportionnel à la qualité de l'individu qu'il représente. Ainsi les meilleurs individus ont plus de chance d'être sélectionné et de participer à l'amélioration de la population.La figure 3.9 présente un exemple de sélection roulette.

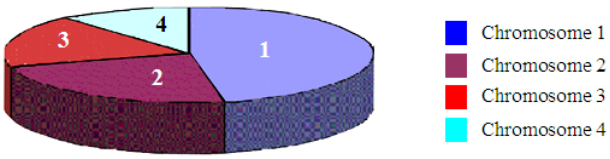


FIG. 3.9 – *Exemple de sélection par roulette*

   - **Sélection sur le rang**: cette méthode de sélection est divisée en deux étapes [117, 118]. Tout d'abord, il faut ranger les individus par ordre croissant (ou décroissant pour un problème de maximisation) de leur qualité. Ensuite, une procédure de sélection similaire à celle de la roulette mais appliquée au rang est utilisée. Cette procédure permet d'attribuer une probabilité de sélection selon leur rang (*"Ranking selection"*).
   - **Sélection par tournoi**: On sélectionne des un sous ensemble de l'ensemble des individus aptes à la reproduction. Ces individus sont sélectionnés de la manière suivante : $\boldsymbol{K}$ individus sont tirés au sort dans la population des individus ($\boldsymbol{K}$ est un paramètre appelé taille du tournoi). Il existe différentes sélections par

tournoi : déterministe ou probabiliste. Dans le cas du tournoi déterministe, le meilleur des $\boldsymbol{K}$ individus gagne le tournoi. Dans le cas probabiliste, chaque individu peut être choisi comme vainqueur avec une probabilité proportionnelle à sa fonction d'évaluation.

2. **Croisement**

   L'opérateur de croisement a pour but d'enrichir la diversité de la population en manipulant les composantes du chromosome. Il consiste à combiner deux individus quelconques (dits parents) avec une probabilité (Px) pour en ressortir deux autres individus (dits enfants) pas forcément meilleurs que les parents.Il existe plusieurs variantes de cet opérateur, mais en général il consiste à couper en un ou plusieurs points deux individus (aux mêmes endroits dans les deux individus) et à échanger les parties situées entre ces points, ces derniers sont générés aléatoirement.La figure 3.10 représente un exemple de croisement en deux points.

FIG. 3.10 – *Fonctionnement des algorithmes génétiques*

3. **Mutation**

   Le rôle de la mutation est de faire apparaître de nouveaux gènes. Il consiste à choisir d'une manière aléatoire un ou plusieurs gènes et à modifier leurs valeurs. Cet opérateur est utilisé avec une probabilité (Pm). Dans un AG binaire, cette probabilité s'effectue sur les gènes en changeant sa valeur de 0 à 1 ou de 1 à 0 et non sur la totalité du chromosome. Mais avec un AG codé autrement, on applique cette probabilité par rapport à l'individu tout entier et non sur les gènes. Si B, généré aléatoirement, appartient à [0,Pm] alors l'opérateur de mutation sera appliqué sur cet individu.

4. **Remplacement**

   Cette dernière étape du processus itératif consiste en l'incorporation des nouvelles solutions dans la population courante. Les nouvelles solutions sont ajoutées à la population courante par remplacement (total ou partiel) des anciennes solutions. Généralement, les meilleures solutions remplacent les plus mauvaises ; il en résulte une amélioration de la population.

Le principe général de fonctionnement des AG est présenté par l'organigramme présenté dans la figure 3.11. L'AG débute par la génération d'une population initiale de N

individus, pour lesquels les valeurs de leurs fonctions objectives sont calculées, les individus seront sélectionnés par une méthode de sélection (roulette, aléatoire, rang ...). Les individus concernés par l'opérateur de croisement seront choisis selon une probabilité Px. Leurs résultats peuvent être mutés par un opérateur de mutation avec une probabilité de mutation Pm. Les individus issus de ces opérateurs génétiques seront insérés par une méthode d'insertion dans la nouvelle population dont la valeur de la fonction objective de chaque individu sera évaluée. Un test d'arrêt sera effectué pour vérifier la qualité des individus obtenus .Si ce test est vérifié alors l'algorithme s'arrête avec une solution optimale, sinon le processus sera réitérer pour la nouvelle population.

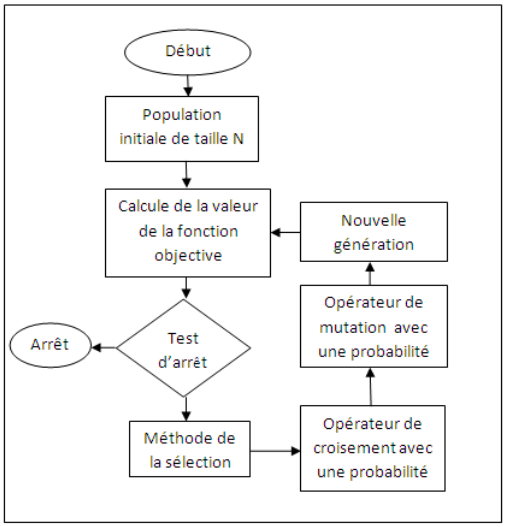


FIG. 3.11 – *Fonctionnement des algorithmes génétiques*

3.3.4 Conclusion

Les algorithmes génétiques sont des algorithmes d'exploration robustes et convergent vers une solution satisfaisante, lorsque leurs paramètres (taille de la population, nombre d'itérations, probabilités de tirages,taux de mutation et de croisement ...) sont choisis d'une manière adéquate. Cependant, le choix de ces derniers ainsi que la fonction d'évaluation et le codage des données est difficile. Il n'y a aucun jeu de paramètres qui est universel pour tous les problèmes considérés du fait que ces valeurs dépendent étroitement du type de problèmes à résoudre. Toutefois, les AG offrent plusieurs avantages comparés aux méthodes conventionnelles tel que:

- **Simple:** ils sont de nature simple, ils utilisent une méthode stochastique afin de se rapprocher de la solution optimale.
- **Efficaces** : ils sont particulièrement efficaces dans des domaines où le calcul de la solution optimale linéairement est très difficile voir impossible.

- **Flexibles :** ils peuvent s'adapter à des configurations et modèles différents, et n'ont pas besoin de paramètres initiaux précis afin de trouver une solution optimale.
- **Robustes :** ils explorent une grande population de solutions, ils peuvent explorer parallèlement différentes sous populations et éviter ainsi de tomber sur un minimum local.

Les AG sont appliqués dans plusieurs domaines tel que le contrôle de systèmes industriels, le traitement d'images, la cryptographie, l'apprentissage des réseaux de neurones, la classification, la sélection d'attributs etc, ainsi que les problèmes d'optimisation où ils se montrent très efficaces.

3.4 Les colonies de fourmis

L'intelligence par essaimes de particules est une nouvelle approche pour la résolution de problèmes. Elle s'inspire du comportement sociale des insectes et d'autre animaux. En particulier, les fourmis artificiels inspirées des colonies de fourmis naturels qui sont beaucoup plus utilisées pour résoudre des problèmes d'optimisation et de classification.

Dans les années 90, des travaux de biologistes s'intéressant de près à la modélisation mathématique et informatique des fourmis ont eu lieu ainsi que l'utilisation concrète de ces modèles. J. L. Deneubourg apparaît comme un pionnier dans le domaine du tri d'objets par des fourmis artificielles [119].

3.4.1 Les fourmis réelles

La fourmi est un insecte social de la superfamille des Formicoides. Elle vit en colonie dans des habitations collectives appelées fourmilières. Les moyens de communication entres les fourmis sont nombreux, toute fois la communication chimique est plus signifiante comparé aux autres. Elle se fait grâce à des phéromones qui sont des substances chimiques olfactives et volatiles.

3.4.1.1 Le fourragement collectif par stigmergie

Les études éthologistes ont montré que dans la nature, les fourmis utilisent la stigmergie pour mener à bien la tâche de recherche de nourriture appelée aussi fourragement. En se déplaçant de leur nid à la recherche d'une source de nourriture, la fourmi laisse sur le chemin qu'elle emprunte une substance chimique de nature volatile appelée « phéromone ». Par ce marquage naturel, elle incite ses congénères à suivre le même trajet. Au début,

les fourmis explorent différents chemins en effectuant des déplacements aléatoires. Une fois qu'un chemin intéressant (menant à une source de nourriture) est découvert, elles y déposent une quantité de phéromone renforçant ainsi son importance et la probabilité d'être choisi par d'autres fourmis de la colonie. D'un autre côté, les mauvais chemins auront tendance à être oubliés voir même disparaître avec l'évaporation de la phéromone. Ce procédé basé sur le mécanisme de rétroaction positive, assure que pendant le fourragement pour la nourriture, les fourmis utilisent la voie d'accès la plus courte car elle sera le plus imprégnée par la phéromone.

3.4.2 Les fourmis artificielles

Une fourmi artificielle est une entité simple dotée d'un comportement similaire ou étendu à celui de la fourmi réelle. Ce comportement doit être élémentaire, restreint et donc facile à programmer. A l'intérieur d'une colonie, les fourmis sont concurrentes et asynchrones, elles coopèrent ensemble pour la résolution de problèmes. Les fourmis artificielles communiquent entre elles indirectement par stigmergie via des modifications de leur environnement (par exemple par dépôt de traces de phéromone artificielle) qui représente la mémoire collective de la colonie.

Chaque fourmi se déplace itérativement d'un état S_i à un état S_j guidée par deux facteurs principaux:

- **Information heuristique:** une mesure de la préférence heuristique pour le déplacement de l'état S_i à l'état S_j. Cette information est connue à priori pour le déroulement de l'algorithme et elle n'est pas modifiée durant l'exécution de ce dernier.
- **Trace des phéromones artificiels:** la mesure du dépôt de phéromone dans les transitions des fourmis allant de l'état S_i vers l'état S_j. Cette information est modifiée durant l'exécution de l'algorithme par les fourmis artificielles.

3.4.2.1 Propriétés des fourmis artificielles

Les principales propriétés associées aux fourmis artificielles sont [120] :

- Chaque fourmi a une mémoire interne utilisée pour la sauvegarde du chemin qu'elle a parcourue.
- Démarrant d'un état initiale " I", chaque fourmi essaye de construire une solution possible a un problème donné.

- Les facteurs de guidage impliqués dans un mouvement de fourmi prennent la forme d'une règle de transition qui est appliquée avant chaque mouvement de l'état S_i à l'état S_j. La règle de transition peut également inclure des contraintes spécifiques au problème et utiliser la mémoire interne des fourmis.
- La quantité de phéromone déposée par chaque fourmi est régie par une règle de mise à jour de phéromone spécifique au problème.
- Les fourmis peuvent déposer des phéromones liés aux états, ou alternativement, aux transitions d'état.
- Le dépôt de phéromone peut se produire à chaque transition d'état pendant la construction de solution.
- Alternativement les fourmis peuvent retracer leurs chemins une fois qu'une solution est construite, et déposer les phéromones le long de leurs chemins individuels.

3.4.3 Principaux algorithmes

Dans la nature, les fourmis arrivent à résoudre différents problèmes d'optimisation liés à leur survie. Cette capacité naturelle a été simulée et transposée pour la résolution de divers problèmes d'optimisation combinatoire. Les premiers travaux dans ce sens sont ceux de Dorigo et al qui ont mimé le comportement collectif de fourragement observé chez les fourmis réelles et l'ont appliqué pour traiter le problème du voyageur du commerce [121] [122].

Ant System (AS) est le premier algorithme ACO (*Ant Colony Optimization*). Il repose sur le comportement de fourragement des fourmis et appliqué pour la résolution du problème du voyageur de commerce (PVC) [123]. La principale caractéristique de AS est que la valeur du phéromone est mise à jour par toutes les fourmis qui ont construit la solution à la même itération.

L'algorithme générale de ACO est présenté dans la figure 3.12

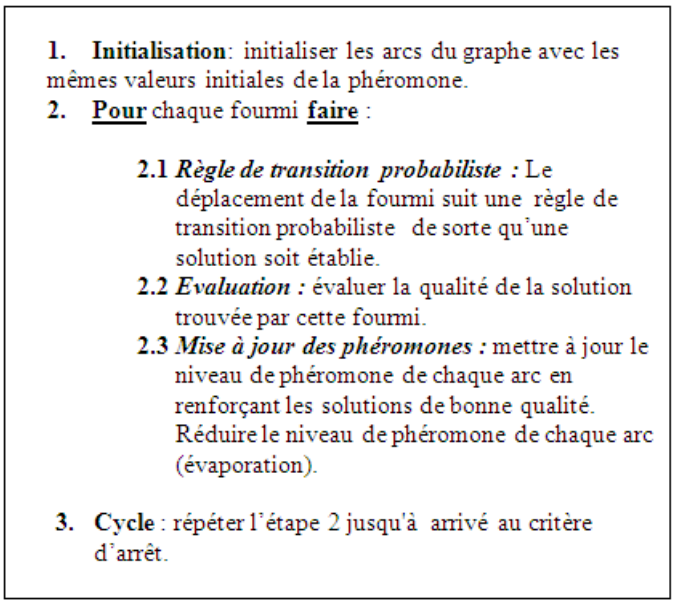


FIG. 3.12 – *Algorithmes générale de ACO*

Le succès de AS a incité de nouveaux chercheurs à l'utiliser en lui ajoutant certaines extensions, d'où les principaux algorithmes suivant:

3.4.3.1 Max–Min Ant System

L'algorithme Max–Min Ant System diffère de AS comme suit [124]:

- Seule la fourmi qui a trouvé la meilleure solution est autorisée à renforcer la phéromone sur tous les arcs constituant son tour.
- La fonction de mise à jour des phéromones est bornée.
- Les valeurs des phéromones sont initialisées à la borne max.

Ceci permet d'éviter que certains chemins d'exploration soient trop favorisés.

3.4.3.2 Ant Colony System

L'algorithme « Ant Colony System » à été introduit par Dorigo et Grambardella en 1996 pour améliorer la performance de AS [125]. ACS se se distingue de AS par les points suivants :

- Le déplacement de la fourmi suit une autre règle de transition dite règle proportionnelle pseudo-aléatoire ;

- Deux méthodes sont utilisées pour la mise à jour des phéromones:
  - Une mise à jour locale qui est effectuée à chaque fin de cycle d'une fourmi.
  - Une mise à jour globale qui est faite une fois que toutes les fourmis ont terminé leurs cycles.Seule la fourmi qui a trouvé la meilleure solution est autorisée à renforcer la phéromone sur tous les arcs constituant son tour.

La mise à jour globale évite de se bloquer dans des solutions sous optimalles (minimums locaux), tandis que la mise à jour locale a pour effet de réduire, de moins en moins, l'interactivité des arcs déjà visités par d'autres fourmis, et donc de favoriser l'émergence des autres solutions que celle déjà trouvées pendant les prochains cycles de l'algorithme.

3.4.3.3 Rank-based Ant System

Rank-based Ant System a été introduit par Bullnheimer [126] .Il incorpore le concept de rang dans la fonction de mise à jour des phéromones comme suit:

- Les fourmis sont rangées par ordre décroissant de la qualité de leurs solutions(langueur du tour).
- La mise à jour des phéromones est réalisée de sorte que seuls les chemins traversés par les n meilleurs fourmis seront mis à jour et la quantité du phéromone déposé dépend du rang de la fourmi.

3.4.4 Conclusion

Les algorithmes de fourmis sont des algorithmes flexibles et robustes, ils peuvent être facilement réutilisables pour des versions modifiées d'un même problèmes et nécessitent peu de changement pour leur application à de nouveaux problèmes d'optimisation combinatoires. Les premières application des ACO été dans le domaine d'optimisation des problème combinatoire NP difficile. Une autre application considérée tôt dans l'histoire des ACO est le routage dans les réseaux, puis ils sont appliqués dans plusieurs domaines notamment, la coloration de graphe, problèmes d'emploi du temps, assignement quadratique ainsi que l'optimisation et la classification où ils ont prouvé leurs succès et etc.

3.5 Conclusion

Dans la littérature il existe une multitude d'algorithmes inspirés de la biologie. Dans ce chapitre nous avons présenté quatre types d'algorithmes à savoir les systèmes immunitaires

artificiels, les réseaux de neurones, les algorithmes génétiques et les colonies de fourmis. Chacun mime une approche biologique différente et a ses propres caractéristiques, ses points forts, ses limites, ses modèles, ses domaines d'application ...etc. Dans le tableau 3.1 nous présentons une comparaison entre ces quatre algorithmes.

L'analyse de ce tableau ressort des points saillants que nous pouvons résumer dans ce qui suit:

Les AIS peuvent être utilisés sous forme de réseaux d'éléments comme les RN, ou sous forme d'ensemble d'éléments travaillant sans notion de communication comme les AG et les ANTs, ce qui augmente leur faculté d'application à différents types de problèmes. Contrairement aux AG, les AIS ont une forte capacité d'apprentissage et de mémorisation ainsi que les réseaux de neurones et les fourmis ce qui est très favorisé dans la classification de données et la reconnaissance de formes.

CARACTERISTIQUES	**AIS**	**RNA**	**AG**	**ANT**
Composantes	Attributs	Neurones	Chromosomes	Fourmis
Apprentissage	Non supervisé	Non supervisé / Supervisé	Pas d'apprentissage	Par renforcement
Mémoire	Cellules mémoires	Les poids des connections	Pas de mémoire	phéromones
Pouvoir explicatif	Fort	Faible	Faible	Fort
Métadynamique	Oui	Non	Oui	Non
Emplacement des composantes	Dynamique	Prédéfinit	Dynamique	Dynamique
Fonction d'évaluation	Fonction de similarité	Fonction d'activation	Fonction fitnesse	Taux de phéromones
Adaptation	Adaptatif	Non adaptatif	Adaptatif	Non adaptatif
Représentation de données	Utilise un codage	utilise un codage	Utilise un codage	Pas de codage
Sauvegarde d'informations	Attributs / Concentration réseaux	Poids des connections	Chromosomes	Phéromone
Evolution	Il évolue	N'évolue pas	Evolue	N'évolue pas
Interaction entre les composantes	Reconnaissance / Connections du réseaux	Connections réseaux	Croisement	Phéromone

Caractéristique	**AIS**	**RNA**	**AG**	**ANT**
Succès d'application	Reconnaissance de forme, détection d'anomalie, classification,	Classification, reconnaissance de forme	Optimisation	optimisation, classification

TAB. 3.1 – ***Tableau comparatif***

Les AIS et les AG sont des systèmes métadynamiques, ils se renouvellent constamment en créant de nouveaux éléments et se débarrassent des éléments les moins utiles, ce qui augmente la qualité des solutions qu'ils génèrent et leur convergence vers une solution optimale, contrairement aux RN qui gardent le même ensemble d'entrées et les colonies de fourmis qui gardent les même fourmis.

Les AIS ont la capacité d'évolution des AG grâce à la sélection clonale, par contre les RN et les ANTs n'ont pas cette faculté et si les données évoluent dans le temps il sera nécessaire de relancer la phase d'apprentissage, ainsi les AIS et les AG s'adaptent aux changements dans les systèmes contrairement des RN et les ANT.

Les RN se comportent comme une boite noir, ils ont un faible pouvoir explicatif des résultats qu'ils génèrent par contre la fonction d'affinité utilisée par les AIS qui se base sur le calcule des distances entre les éléments explique bien les résultats générés.

Nous constatons que chacun des algorithmes cités dans le tableau 3.1 a ses propres caractéristiques et nous constatons que les AIS en plus de leurs propres caractéristiques ils incluent d'autres qui sont spécifiques aux réseaux de neurones, les AG et les ANTs, ce qui augmente leur succès d'application dans plusieurs domaines. Ils sont autodynamique robuste et ont une bonne faculté d'apprentissage et de mémorisation, ce qui est très favorisé dans le domaine de reconnaissance de forme et de classification desquels dérive notre application, "'la reconnaissance de l'écriture manuscrite arabe"', pour cela nous choisissons les AIS pour les utiliser comme classifieur dans notre application. Pour avoir des résultats plus performants, la tendance est de construire des systèmes hybrides qui combinent entre plusieurs algorithmes. Les AG ont des capacités d'adaptation et d'évolution, et sont réputés d'être des bons optimiseurs pour cela nous choisissons cette classe d'algorithmes pour l'optimisation des résultats de classification des AIS. Notre proposition hybride sera détaillée dans le chapitre prochain.

Chapitre 4

Contribution à la reconnaissance de l'écriture arabe manuscrite

Plusieurs travaux traitant la reconnaissance de l'écriture arabe ont été réalisés mais le taux de reconnaissance reste toujours insatisfaisant. Ce dernier dépend du classifieur et des attributs de classification utilisés. Pour avoir un taux de reconnaissance élevé, l'idée est de concevoir un algorithme hybride bio-inspiré pour tirer avantages de leurs efficacités. Dans ce chapitre nous présentons notre algorithme hybride immuno-genetic, qui se base sur les systèmes immunitaires artificiels et les algorithmes génétiques. Pour la reconnaissance des mots arabe, nous avons utilisé les systèmes immunitaires artificiels qui ont la faculté de mémorisation et de bon apprentissage, de plus ils sont bien utilisés dans le domaine de reconnaissance de formes, puis pour la sélection de la liste d'attributs de classification qui donne le meilleur taux de reconnaissance nous avons choisi d'utiliser les algorithmes génétique, qui sont réputés être de bons optimiseurs et bien appliqués dans ce genre d'application.

4.1 Contribution

Dans cette section nous présentons notre contribution à la reconnaissance de l'écriture arabe manuscrite, notre algorithme hybride bio-inspiré ainsi que la liste des attributs de classification.

4.1.1 Proposition des attributs

Dans la littérature, il existe plusieurs types d'attributs utilisés dans la classification et la reconnaissance de l'écriture arabe à savoir les caractéristiques structurelles, statistiques,

basés sur les transformés ... qui sont détaillés dans la section 2.3.3.3 du chapitre 2. Dans notre approche nous utilisons les caractéristiques structurelles d'un mot qui décrivent bien l'écriture arabe manuscrite. La figure 4.1.1 présente une phrase écrite en arabe avec des caractéristiques structurelles. Nous proposons une liste de 18 attributs(caractéristiques) présentés dans le tableau 4.1, ces attributs sont calculés à partir des images squelettisées des mots à reconnaître.

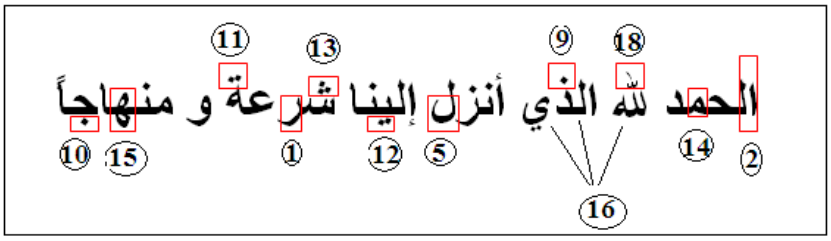


FIG. 4.1 – *Une phrase écrite en arabe*

Attributs de classification		
1	PJ	Présence de Jambes
2	PH	Présence de hampe
3	PB	Présence de boucle
4	PPD	Présence de points diacritique
5	PC	Présence de concavités
6	NbCV	Nombre de concavités
7	NbH	Nombre de Hampes
8	NbJ	Nombre de Jambes
9	NbPDUH	Nombre de points diacritique unique haut
10	NbPDUB	Nombre de points diacritique unique bas
11	NbPDDH	Nombre de points diacritique double haut
12	NbPDDB	Nombre de points diacritique double bas
13	NPDT	Nombre de points diacritique triple
14	NbBU	Nombre de boucle unique
15	NbBD	Nombre de boucle double
16	NbSM	Nombre de sous mots
17	NbC	Nombre de caractères
18	Pcc	présence de chadda

TAB. 4.1 – ***Les attributs de classification***

4.1.2 Classification

Il existe une multitude de méthodes de classification de données qui dépendent du domaine d'application du problème à traiter. La reconnaissance de l'écriture manuscrite arabe nécessite un apprentissage et relève du domaine de la reconnaissance de formes d'où nous proposons d'utiliser comme classifieur les systèmes immunitaire artificiels, qui se distinguent des autres méthodes de classification en particulier les méthodes bio-inspirés par leur faculté de mémorisation, de reconnaissance, d'auto-organisation , d'apprentissage et leur bonne application dans le domaine de reconnaissance de formes. Par rapport à d'autres algorithmes bio-inspirés, les AIS offrent des modèles et des algorithmes différents (la sélection clonale, sélection négative, réseaux immunitaires) conçus pour différentes applications. Dans notre approche nous nous intéressons à la sélection clonale qui s'applique bien pour la reconnaissance de forme de laquel dérive notre application.Les anticorps du systèmes sont des mots arabe appris par le classifieur, les antigènes sont les mots arabe à reconnaître, ils sont représentés par des vecteurs de caractéristiques, ces dernières sont choisies par l'algorithme génétique. Les anticorps et les cellules mémoire sont représentés par des vecteurs de même type que ceux des antigènes. L'affinité entre les anticorps et les antigènes est calculée avec une distance correspondante aux données traitées. L'antigène est affilié à la classe avec laquelle il aura la valeur d'affinité la plus petite.

4.1.3 Sélection des attributs

La sélection d'attributs consiste à rechercher un sous-ensemble d'attributs pour améliorer la classification et réduire le temps de traitement. Il existe deux types de méthodes de sélection d'attributs : les méthodes enveloppantes *(wrapper)* qui se servent d'un classifieur pour évaluer le sous-ensemble d'attributs sélectionnés et les méthodes filtrantes où l'évaluation se fait indépendamment du classifieur.

Dans notre approche, nous utilisons les caractéristiques structurelles d'un mot qui sont nombreuses. Nous proposons d'utiliser un maximum d'attributs structurels pour la classification, cependant, la recherche exhaustive de la combinaison de caractéristiques qui génère un taux de classification élevé est très coûteuse en temps de calcule. Pour remédier à cette lacune, nous avons choisi d'utiliser les algorithmes génétiques pour la recherche de cette combinaison, qui s'appliquent bien sur des problèmes d'optimisation et la sélection d'attributs de classification en générale. La population initiale des AG est constituée d'un ensemble de vecteurs(chromosomes) représentants des combinaisons d'attributs, le codage des vecteurs est binaire et la fonction fitness est calculée à partir du taux de classification pour un vecteur d'attributs donné.

4.1.4 Architecture de la proposition

Notre contribution pour la reconnaissance de l'écriture arabe se base sur l'hybridation de deux algorithmes bio-inspirés les systèmes immunitaires artificiels et les algorithmes génétiques. Nous avons proposé une liste de caractéristiques structurelles des mots arabe qui serons utilisés comme attributs de classification pour le classifieur immunitaire, cette liste sera raffinée par les algorithmes génétiques de manière à avoir à la fin une liste contenant des attributs qui donne le meilleur taux de reconnaissance. Le fonctionnement générale de notre algorithme est illustré dans le schéma de la figure 4.2. L'AG reçoit en entrée un ensemble de vecteurs d'attributs avec un codage binaire, 1 signifie la prise en compte de l'attribut correspondant et 0 pour le cas contraire. A chaque itération l'AG sélectionne un vecteur V de la population, puis l'algorithme de classification immunitaire sera lancé avec les attributs correspondants au vecteur sélectionné précédemment et qui ont leurs valeurs à 1, le taux de reconnaissance pour cet ensemble d'attributs sera la valeur de la fonction fitness pour ce vecteur V dans l'AG. Ce procédé est réitéré jusqu'à ce que le critère d'arrêt de l'AG soit atteint (nombre d'itération). A la fin de l'algorithme on aura en sortie le meilleur taux de reconnaissance qui correspond à la combinaison d'attributs optimale.

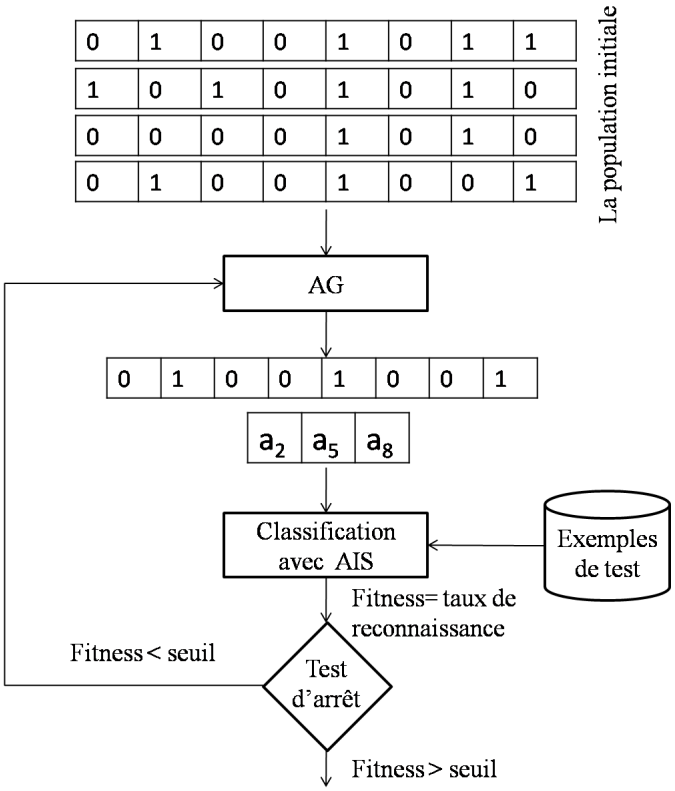


FIG. 4.2 – *Algorithme immino-genetic*

4.2 Implémentation

Dans cette section nous présentons les moyens et les méthodes techniques que nous avons utilisé pour implémenter notre solution proposée ci-dessus. Nous présentons les outils de développement et des tests, la structure de notre application et ses différentes fonctionnalités(calcul des attributs, classification, optimisation).

4.2.1 Outils de développement

Pour l'implémentation et le test de notre algorithme immino-genetic, nous avons utilisé l'environnement Windows, et le langage de programmation java avec l'IDE éclipse. L'implémentation du classifieur immunitaire est réalisé avec l'API weka qui offre un package riche pour les systèmes immunitaires artificiels weka.classifiers.immune. Le Framework JGAP (java genetic algorithm programming) est utilisé pour l'optimisation génétique. L'ensemble des testes s'est porté sur un sous ensemble de la base des mots manuscrits des villes tunisiennes IFN /ENIT.

1. **weka**

   Weka(Waikato Environment for Knowledge Analysis)[1] est une boite à outils complète pour l'apprentissage automatique et la fouille de données développée avec java à l'université de Waikato, New Zealand, Weka est un logiciel gratuit disponible sous la licence GNU(General Public License).

   Weka supporte différentes tâches d'analyse de données, prétraitement, segmentation, classification, régression et visualisation. Cette API peut être utilisée via son interface graphique ou en lignes de commandes, comme on peut faire appelle à ses différentes classes à partir de nos programme java.

   Nous avons choisi d'implémenter notre algorithme avec Weka pour ses multiples avantages:

   - Gratuit et disponible.
   - Offre un package complet pour les Systèmes immunitaires .
   - Utilise un format simple pour la représentation de données(fichier ARFF).
   - Implémenté en java, un langage orienté objet très puissant et gratuit.

2. **JGAP**

   JGAP (Java Genetic Algorithms Package), est un framework java pour la mise en oeuvre des algorithmes génétiques. Il offre les moyens techniques et algorithmiques nécessaire pour appliquer les algorithmes génétiques aux problèmes d'optimisation.

1. www.cs.waikato.ac.nz/ml/weka

Les éléments de l'algorithme génétique sont implémentés dans des classes java que nous pouvons utiliser directement à partir de nos programmes. Nous avons utilisé ces classes pour implémenter l'optimiseur génétique de notre solution.

3. **IFN/ENIT**

   IFN/INIT est une base d'images pour l'évaluation des OCRs de la langue arabe développée à l'école nationale d'ingénieur de tunis (ENIT) en collaboration avec l'université Braunschweig, Almagne (Institute for Communications Technology (IFN)). Cette base contient des noms manuscrits des villes et villages tunisiens, elle est construite à partir de 2200 pages de formulaire remplis par un ensemble de 411 scripteurs différents. Les images de la base sont réparties dans 4 sous ensembles(a-e) avec deux formats différents, images bmp et tif. Pour chaque nom de ville ou village, les informations de base sont codées dans un fichier .TRU, par exemple la taille de l'image, le nom de la ville en codage de la langue arabe, le nombre de caractères et etc. La figure4.3 illustre un exemple d'un fichier au format TRU.

```
01: COM: IFN/ENIT-database truth (label) file
02: COM: http://www.ifnenit.com
03: COM: IfN, TU-BS
04: COM: di45_019.tif coming from pb377_6.tif
05: X_Y: 498 87
06: BDR: begin data record
07: LBL: ZIP:3032;AW1:مركزدرويش;AW2:maB|raE|keB|zaE|daA|raA|waA|yaB|shE|;QUA:YB1;ADD:P6
08: CHA: 9
09: BLN: 56,42
10: EDR: end of data record
```

FIG. 4.3 – *Exemple de fichier .TRU*

- Lignes 0-1: sont des commentaires.
- Ligne 05 : taille de l'image en pixel.
- Ligne 06 : début des informations utiles dans le fichier.
- Ligne 07 : contient les informations du vrai nom de ville,elle inclut:
  - Zip: code postale tunisien.
  - AW1: nom de la ville avec le codage "Arabic Windows".
  - AW2: codage latin des caractère du nom de la ville.
  - QUA: qualité des borne de la ligne de base(B1=:OK;B2=:bad).
  - ADD: nombre de sous mots dans le nom de la ville.
- Ligne 08 : nombre de caractères.
- Ligne 09 : coordonnés de la ligne de base dans l'axe Y.
- Ligne 10 : indique la fin des informations.

4.2.2 Organisation de l'application

Le bute de notre application est de créer un fichier au format "arff", qui sera utilisé par le classifieur immunitaire pour la reconnaissance des noms de villes de la base IFN/ENIT, puis optimiser les résultats de classification en les passant aux algorithmes génétiques qui sélectionnent une liste d'attributs(caractéristiques) qui génère un taux de reconnaissance élevé. Pour cela nous avons organisé notre application sous fourme de classes java dont le diagramme de classe simplifié est présenté dans la figure 4.4. Les images des noms des villes de la base "IFN/ENIT" sont de nature prétraitées (binarisées, lissées et redressées). Pour faciliter l'extraction des caractéristiques(attributs de classification) nous procédons d'abord à la squelettisation des images des noms de villes et cela avec la classe *Thining.java* qui implémente l'algorithme de squelettisation de hildith. La classe *IfnPath.java* retourne le chemin d'accès au dossier contenant les images des noms des viles ainsi que les fichiers "'TRU"' ce chemin sera utilisé par les classe *ArffManager.java*. Dans la classe *FeaturExtraction.java* nous extrayons les attributs de classification. Cette classe retourne un vecteur de caractéristiques de taille égale au nombre d'attributs de classification et elle est instantiée pour chaque nom de ville de la base des images. Ce vecteur retourné sera utilisé par la classe *ArffManager.java* pour la création d'un fichier au format "arff" qui sera utilisé par la classe *ImmunClassifier.java* pour la reconnaissance des noms des villes. Le taux de reconnaissance retourné par la classe *ImmunClassifier.java* sera comme entrée pour la classe *MyFitness.java*, qui restitue le taux de reconnaissance et cela suivant le classifieur sur lequel est basée la classification(CLONALG ou AIRS) et l'associe à l'individu correspondant. Puis les résultats serons passés à la classe *GeneticSelector.java* pour la sélection de la combinaison d'attributs de classification candidate. La liste des attributs qui sera généré par cette classe va âtre retourner à la classe *ImmunClassifier.java* pour l'évaluation et ainsi de suite jusqu'a rencontre du critère d'arrêt.

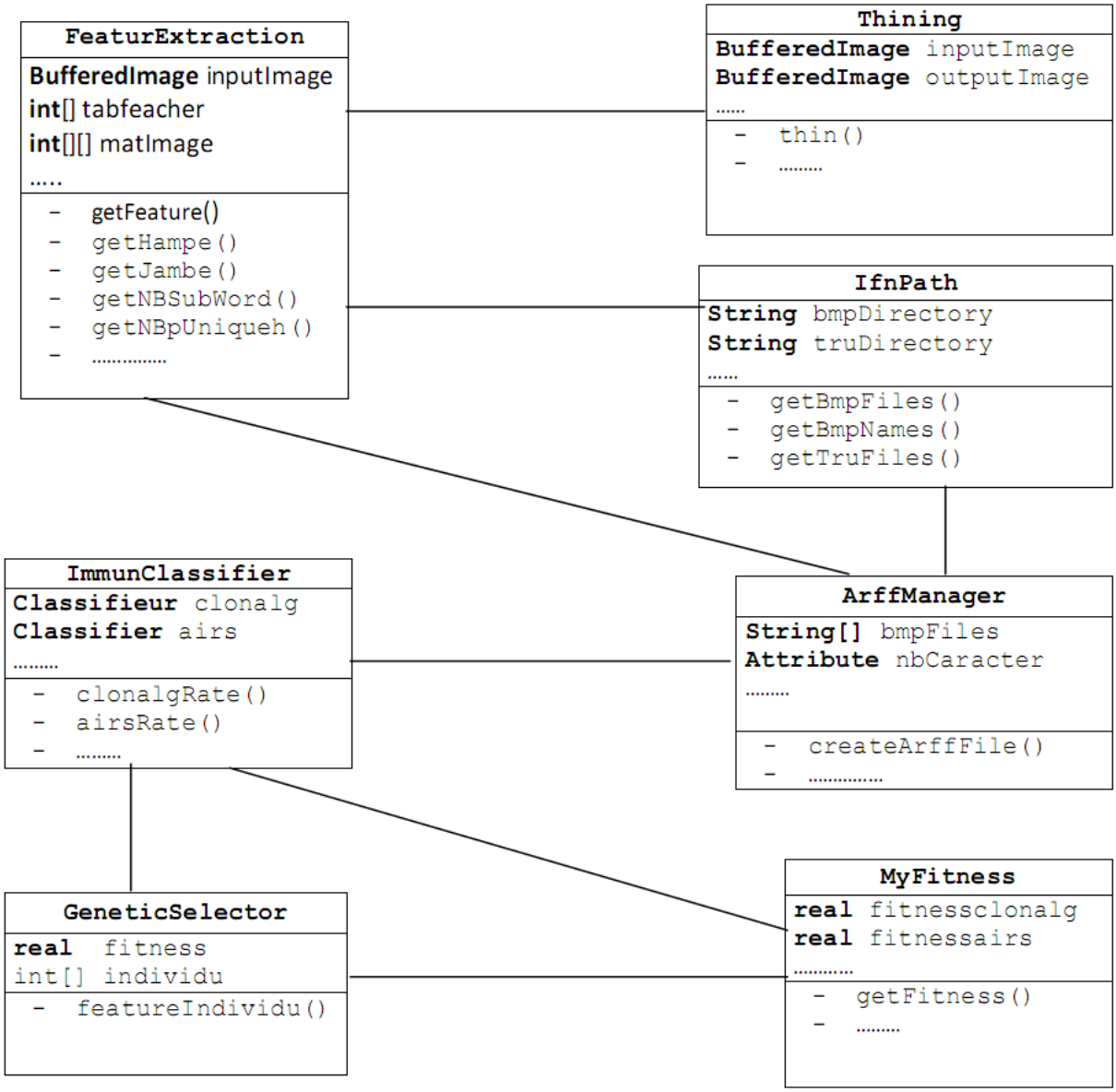


FIG. 4.4 – *Diagramme de classes*

4.2.2.1 Calcul des attributs

Nous proposons de squelettiser les mots(noms de villes) à reconnaître pour faciliter l'extraction de leurs caractéristiques proposées précédemment qui serons utilisées comme attributs de classification par notre classifieur. Le calcule de ces attributs est réalisé dans la classe *FeaturExtraction.java*. L'ensemble des calcules se base sur les matrices des images des mots à reconnaître calculées avec la méthode *getMatImage()*. Comme exemples de méthodes de calcules des attributs, la méthode *getNumHamps()* qui calcule le nombre de hampes dans un mot , l'attribut nombre nombre de sous mots est calculé avec la méthode *getNumSubWord()*, cette dernière se sert de la méthode *getVerticalHistog()* dans ses cal-

cules. La méthode *getFeature()* retourne un tableaux de type entier qui contient les valeurs des attributs de classification retournées par chacune des méthodes correspondante. Ce vecteur sera utilisé pour la création du fichier .arff avec la méthode *createArffFile()* de la classe *Arffmanager.java*.

4.2.2.2 Classification

Pour la reconnaissance des mots (noms de villes), nous utilisons la classe *ImmunClassifieur.java*. Cette classe utilise le fichier arff généré avec la méthode *createArffFile()* de la classe *Arffmanager.java*. Nous proposons de faire appelle à deux classifieurs immunitaires différents implémentés dans weka et cela pour comparer les résultats de classification qu'ils grenèrent. La méthode *clonalgRate()* instancie la classe "CLONALG.java" qui est une implémentation de l'algorithme de la sélection clonale immunitaire. La méthode *airsRate()* instancie la classe *AIRS1.java* qui est une implémentation d'un algorithme immunitaire dédié à la classification. Chacune de ces deux méthodes essaie de classer les noms des villes dans la classe de la ville à laquelle elle ont la valeur d'affinité la plus petite, et retourne à la fin du processus de classification un taux de reconnaissance qui sera utilisé par la classe *GeneticSelector.java*.

4.2.2.3 Sélection des attributs

Pour la sélection d'une combinaison d'attributs de classification générant le meilleur taux de reconnaissance , nous utilisons la classe *GeneticSelector.java*. La fonction d'évaluation des individus est implémenté dans la classe *MyFitness.java*, qui hérite de la classe *FitnessFunction.java* du package JGAP. La valeur de la fonction fitness des individus sera restitué à partir des méthodes *airsRate()* et *clonalgRate()* pour chacun des deux classifieurs et cela en utilisant la méthode *getFitness()*. Puis les résultats serons passés à la classe *GeneticSelector.java*. A la fin du processus de sélection l'individu élu par la méthode *featureIndividu()* de la classe *GeneticSelector.java* sera en entrée pour la classe *ImmunClassifieur.java* pour réévaluer et reclasser l'ensemble des données suivant l'ensemble des attributs pris en compte (qui ont la valeur 1 dans le vecteur d'attributs) par le sélecteur génétique.

4.3 Résultats et discutions

Pour l'évaluation des performance de notre algorithme proposé "immino-genetic", nous avons utilisé un sous ensemble de la base des ville IFN/ENIT. Nous avons utilisé les sous

ensembles a, b, c pour l'apprentissage et l'ensemble d pour le test. Nous avons utilisé deux modèles de classifieurs immunitaire implémenté dans Weka, AIRS (Artificial immun recognition system) et CLONALG et ceci afin de comparer les taux de reconnaissance généré par ces derniers et choisir celui qui génère le meilleur taux. Nous avons utilisé l'algorithme génétique pour la sélection d'une combinaisons d'attributs générant le meilleur taux de classification.

La liste des attributs de classification proposés est présenté par un vecteur à valeurs binaires de 1 et/ou 0. Le 1 signifie la prise en compte de l'attribut par le classifieur et 0 signifie que l'attribut n'est pas pris en compte. L'ordre des attributs de classification sont présentés dans le tableaux 4.1. Le vecteur qui a toutes ces valeurs à 1 veut dire que tous les attributs de classification sont pris en compte.

Dans le tableaux 4.2 nous avons présenté quelques taux de reconnaissance généré par les algorithmes AIRS et CLONALG en utilisant quelques combinaisons de d'attributs, la liste exhaustives des combinaisons d'attributs est très grande elle est égale à 2^{18} combinaisons. Pour cela nous exposons dans le tableaux 4.3 le meilleur taux de reconnaissance généré par notre algorithme ainsi que la combinaison d'attributs de classification correspondante.

La liste des attributs	**Taux de reconnaissance**	
	AIR	**CLONALG**
1111111111111111	65%	64 ,5%
1101001110111011	66%	65 ,34%
1101010101011110	67.74%	66%
1101100101110111	61 ,3%	60,5%
0011011011110111	66 ,62%	65 ,77%
1110111011000111	63%	62%
1100100011011111	61 ,66%	60 ,4%
0011010100111010	62,4%	61 ,4%

TAB. 4.2 – *Résultat de classification avec AIRS et CLONALG*

La liste des attributs	Taux de reconnaissance	
	IMINO-GENETIC (A base de AIRS)	IMINI-GENETIC (A base de CLONALG)
1111111111111111	65%	64 ,5%
0000001111111010110	70%	69,2%

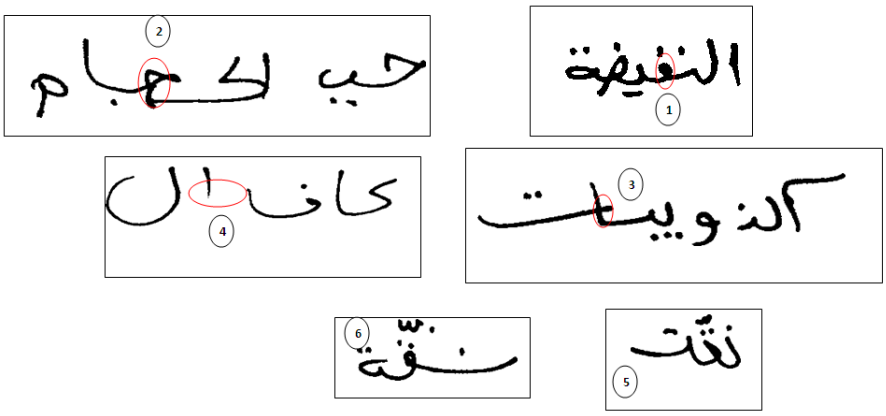
TAB. 4.3 – ***Résultat de classification IMMINO-GENETIC***

L'analyse des deux tableaux 4.2 et 4.3 montre que le taux de reconnaissance dépend du type de classifieur et des attributs de classification utilisés. Les résultats présentés dans le tableaux 4.2 montre que les taux varient d'une combinaison à une autre et d'un classifieur à un autre. L'algorithmes AIRS génère des taux de reconnaissance plus élevés que ceux générés par CLONALG ceci peut être expliqué par le fait que AIRS est un algorithmes immunitaire dédié à la classification(reconnaissance).

Les résultats présentés dans le tableaux 4.3 montrent que le taux de reconnaissance est amélioré en utilisant notre algorithme et il est meilleur en utilisant AIRS comme classifieur et la combinaisons d'attributs suivante: NbH, NbJ, NbPDUH, NbPDUB, NbPDDH, NbPPDB, NbBU, NbSM, Nbc a donnée le meilleur taux de reconnaissance par rapport aux résultats générés en utilisant la totalité des attributs et cela peut être justifié par le fait que quelques attributs peuvent freiner la classification.

Les taux de reconnaissance générés sont satisfaisant par rapport à la complexité du manuscrit de la langue arabe tel que la forte cursivité, la variabilité de l'écriture manuscrite d'un scripteur à un autre et autre ...etc.

La figure 4.4 est montre quelques problèmes du manuscrit arabe



TAB. 4.4 – *Quelques mots manuscrit arabes*

1. La lettre encerclée dans le mot 1 doit être écrite comme une boucle, alors qu'ici elle est remplie.
2. La lettre dans encerclée dans le deuxième mot doit être ouverte et détecté comme concavité alors qu'ici elle sera détecté comme boucle.
3. Les lettres encerclées dans le troisième mot doivent être séparé et détecté comme sous mots différent alors qu'ici ils sont connectés et serons pris pour un seul mots.
4. Les lettres encerclées dans le troisième mot doivent être connecté alors qu'ici sont séparé et seront pris pour deux sous mots.
5. Le cinquième et le sixième mot sont les même alors qu'ils sont écrits de deux manières différente.

4.3.1 Conclusion

Dans ce chapitre nous avons présenté notre contribution à la reconnaissance de l'écriture manuscrite, l'algorithme hybride immino-genetic, ainsi qu'une liste d'attributs de classification. Nous avons présenté l'architecture de la proposition, les outils utilisés pour son implémentation et teste et en fin nous avons discuté les résultats obtenues.

Conclusion et perspectives

Dans ce mémoire nous avons proposé un algorithme hybride bio-inspiré pour la reconnaissance de l'écriture manuscrite arabe. Pour cela nous avons fait une étude sur les systèmes de reconnaissance d'écriture en générale et arabe en particulier ainsi que les techniques de classification bio-inspirées du datamining.

Plusieurs travaux traitant cette problématique ont été réalisé mais ils révèlent une insuffisance envers la complexité de l'écriture arabe(cursivité, points diacritiques et etc) en générale et de son manuscrit en particulier qui est dû au changement de l'écriture inter et intra scripteurs. Pour pallier à ça, l'idée est de combiner entre les algorithmes bio-inpirés pour profiter des avantages qu'ils offrent. Nous avons proposé d'hybrider entre les systèmes immunitaires artificiels et les algorithmes génétiques. Les systèmes immunitaires utilisés comme outil de classification pour la reconnaissance des mots arabe manuscrits et les algorithmes génétiques comme optimiseur appliqués sur l'ensemble des attributs caractérisant les mots à reconnaître(caractéristiques structurelles) pour la sélection d'une liste d'attributs pertinente et qui génèrent un taux de reconnaissance élevé.

Nous avons appliqué notre algorithme pour la reconnaissance des mots à savoir l'ensemble des villes tunisiennes IFN/ENIT et les résultats sont satisfaisant par rapport à la complexité du manuscrit arabe. Cette application peut présenter beaucoup d'intérêts dans le tri automatique du courrier et la reconnaissance des montants des chèques.

Perspectives:

Nos perspectives sont axées essentiellement sur les points suivants :

- Valider l'algorithmes que nous avons proposé en l'appliquant sur un texte arabe manuscrit quelconque.
- Introduire l'aspect sémantique à travers un dictionnaire ou une ontologie.

Bibliographie

[1] Oded Maimon, Lior Rokach, Data Mining and Knowledge Discovery Handbook (Second Edition), Springer, ISBN 978-0-387-09822-7, e-ISBN 978-0-387-09823-4, 2010.

[2] Weiss, S. I., and Kulikowski, C. 1991. "'Computer Systems That Learn: Classification and Prediction Methods from Statistics"', Neural Networks, Machine Learning, and Expert Systems. San Francisco, Calif.:Morgan Kaufmann.

[3] Jain, A. K., and Dubes, R. C. 1988. "'Algorithms for Clustering Data. Englewood Cliffs, N.J.: Prentice- Hall.

[4] S. Tufféry. Data mining et statistique décisionnelle. Editions Technip, 2007.

[5] A. R. Webb. Statistical pattern recognition. John Wiley Sons (2nd edition), 2002.

[6] U. Fayyad, G.Piatetsky Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases"','American Association for Artificial Intelligence, 1996 .

[7] Antoine Cornuéjols, Laurent Miclet and Yves Kodratoff,"'Apprentissage artificiel Concepts et algorithmes"', Groupe Eyrolles, 2002, ISBN : 2-212-11020-0.

[8] Han, J. et Kamber, M. Data Mining : Concepts and Techniques. Morgan Kaufman,2001.

[9] H. Liu, L. Huan, and H. Motoda. Feature selection for knowledge discovery and data mining. Springer, 1998.

[10] C. M. Bishop. Neural networks for pattern recognition. Oxford University Press, 1995.

[11] H. Liu and L. Yu. Feature selection for data mining. Technical report, Arizona State University, 2002.

[12] R. Kohavi and G. H. John. Wrappers for feature subset selection. Artificial Intelligence, 97(1-2) :273-324, 1997.

[13] S. Theodoridis and K. Koutroumbas. Pattern Recognition.Academic Press, 2006.

[14] [Liu and Yu, 2005] H. Liu and L. Yu. Toward integrating feature selection algorithms for classication and clustering. IEEE Transactions on Knowledge and Data Engineering, 17(4) :491-502, 2005.

[15] Y. Bennani. Sélection de variables. Revue d'intelligence Artificielle, 15(3-4) :303-316, 2001.

[16] M. Kudo and J. Sklansky. Comparison of algorithms that select features for pattern classifiers. Pattern Recognition, 33(1) :25-41, 2000.

[17] A. Cornuéjols and L. Miclet. Apprentissage artificiel : concepts en algorithmes. Eyrolles, 2002.

[18] A. Nazif, A system for the recognition of the printed Arabic characters, Master's Thesis, Faculty of Engineering, Cairo University, 1975.

[19] N. Ben Amara ,« Utilisation des modèles de Markov cachés planaires en reconnaissance de l'écriture arabe imprimée ». Thèse de doctorat,Tunis, 1999

[20] Fahmy, S.Al Ali : «Automatic recognition of handwritten Arabic characters using their geometrical features ». Studies in informatics and control journal (SIC journal), vol. 10, No 2, 2001

[21] B. Al-Badr, S.A. Mahmoud : «Survey and bibliography of Arabic optical text recognition ». Signal processing, vol. 41,elsevier, 1995.

[22] L. M. Lorigo, V. Govindaraju, Offline Arabic Handwriting Recognition:A Survey,ieee transactions on pattern analysis and machine intelligence, vol. 28, no. 5, may 2006

[23] S. Madhvanath , V. Govindaraju "'The Role of Holistic Paradigms in Handwritten Word Recognition"',ieee transactions on pattern analysis and machine intelligence, vol. 23, no. 2, february 2001

[24] A. Ali Aburas et M. E. Gumah,"'Arabic Handwriting Recognition: Challenges and Solutions"' ,IEEE ,2008.

[25] A. Hamid, R. Haraty, A Neuro-heuristic Approach for Segmenting Handwritten Arabic Text, ACS/IEEE International Conference on Computer Systems and Applications (AICCSA 2001), Lebanon, 2001.

[26] R. Haraty and C. Ghaddar, "Arabic Text Recognition," Int'l Arab J.Information Technology, vol. 1, 2004.

[27] L. Souici-Meslati and M. Sellami, "A Hybrid Approach for Arabic Literal Amounts Recognition," The Arabian J. Science and Eng.,vol. 29, 2004.

[28] A.Amin, Recognition of printed arabic text based on global features and decision tree learning techniques, Pattern Recognition Society. Published by Elsevier Science,2000.

[29] R. Safabakhsh and P. AdibiNastaaligh , handwritten word recognition using a continuous-density variable-duration HMM, the Arabian Journal for Science and Engineering, Volume 30, Number 1B, 2005.

[30] A. Amin et G. Masini, "Machine recognition of multifont printed Arabic texts", Proc. 8th Internat. Joint Conf, on Pattern Recognition, Paris, France, October 1986.

[31] B. Parhami and M. Taraghi, “Automatic recognition of printed Farsi texts”, Pattern Recognition, Vol. 14, NO. 1, 1981,

[32] Ethnologue: Languages of the World, 14th ed. SIL Int’l, 2000.

[33] L.Zheng and al, A new algorithm for machine printed Arabic character segmentation,Pattern Recognition Letters 25 (2004).

[34] A.M. Zeki and M. S. Zakaria , Challenges in Recognizing Arabic Characters,17th natinal conference in informatic, saoudi arabia,2004.

[35] B. Timsari and H. Fahimi, Morphological Approach to Character Recognition in Machine-Printed Persian Words, Proceeding of SPIE, Document Recognition III, San Jose, CA,1996.

[36] A. Nouh and al , ”A proposed algorithm for thinning binary Arabic character patterns”, Proc. 1st Kuwaiti Computer Conf, Kuwait, March 1989.

[37] N.J. Naccache and R. Shinghal, ”SPTA: A proposed algorithm for thinning binary patterns”, IEEE Trans Systems Man Cybernet., Vol. 14, No. 3, May 1984,

[38] L. Lam, S. Lee and C. Suen, ”Thinning methodologies - A comprehensive survey”, IEEE Trans. Pattern Anal Machine lntell., Vol. 14, No. 9, September 1992,

[39] Zainodin, I., D. Khairuddin, and S. Horani. 1994. Sequential thinning of binary images. Sains Malaysia.

[40] Jang, B., and R. T. Chin. One-pass parallel thinning analysis, properties, and quantitative evaluation. IEEE Trans. On Pattern Analysis and Machine Intell.1992

[41] A. AL-Shatnawi and K. Omar,” A Comparative Study between Methods of Arabic Baseline Detection”, International Conference on Electrical Engineering and Informatics Selangor, Malaysia,IEEE, 2009

[42] T. Steinherz, E. Rivlin, N. Intrator: «Off-line cursive word recognition: a survey ». International journal on document analysis and recognition, 1999.

[43] A. Cheung, M. Bennamoun and N. W. Bergmann, ”‘An Arabic Optical Character Recognition System using Recognition-Based Segmentation”’, Pattern Recognition, 2001.

[44] A.M. Zeki,”’The Segmentation Problem in Arabic Character Recognition The State Of The Art”’,IEEE,2005.

[45] B. Parhami and M. Taraghi, Automatic Recognition of Printed Farsi Texts, Pattern Recognition, 1981.

[46] A. Amin and Masini G., Machine Recognition of Arabic Cursive Words, SPIE 26th International Symposioum on Instrument Display, Application of Digital Image Processing IV, Vol. 359, San Diego, Aug 1982.

[47] Amin, A., Recognition of Arabic hand-printed mathematical formulae. Arabian J. Sci. Eng. 16 (4B), 1991.

[48] Altuwaijri, M.M., Bagoumi, M.A., 1994. Arabic text recognition using neural networks. In: IEEE Internat. Symp. on Circuits and Systems (ISCAS 94). May 1994, vol. 6,.

[49] B. M. Kurdy and A. Joukhadar, Multifont Recognition System for Arabic Characters, 3rd Intemational Conference and Exhibition on Multi-lingual Computing (Arabic and Roman Script), University of Durham, UK, 1992.

[50] M. B. Kurdy and M. M. AlSabbagh, Omnifont Arabic Optical Character Recognition System, International Conference on Information and Communication Technologies: from Theory to Applications, Damascus, Syria, 2004.

[51] R. Azmi and E. Kabir, A New Segmentation Technique for Omni-font Farsi Text, Pattern Recognition Letters, Vol. 22, 2001.

[52] F. El-Khaly and M. A. Sid-Ahmed, Machine Recognition of Optically Captured Machine Printed Arabic Text, Pattern Recognition, 1990.

[53] K. Jambi, Design and Implementation of a System for Recognizing Arabic Handwritten Words with Learning Ability, Ph.D. Thesis, Illinois Institute of Technology, Chicago, 1991.

[54] Y. Al-Ohali, Development and Evaluation Environment for Typewritten Arabic Character Recognition, M.Sc. Thesis, King Saud University, Riyadh, 1995.

[55] T. Sari and al , Off-line Handwritten Arabic Character Segmentation and Recognition System: ACSA, Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR'02),IEEE, 2002.

[56] C. Olivierand al , Segmentation and Coding of Arabic Handwritten Words, International Conference on Pattern Recognition (ICPR '96), Vienna, Austria, Vol. 3,1996.

[57] S. Wshah and al,Segmentation of Arabic Handwriting based on both Contour and Skeleton Segmentation,2009 10th International Conference on Document Analysis and Recognition,IEEE,2009

[58] B.M.F. Bushofa, M. Spann, Segmentation and recognition of Arabic characters by structural classification,Image and Vision Computing 15 , Elsevier,1997.

[59] A. Belaïd, Analyse et reconnaissance de documents, Cours INRIA: le Traitement électronique de Documents, Collection ADBS, 3-7 octobre, Aix-en-Provence, 2002.

[60] B. Al-Badr R.M. Haralick, Segmentation-Free Word Recognition with Application to Arabic,IEEE,1995

[61] J. H .AIKhateeb and al , Word-based Handwritten Arabic Scripts Recognition using DCT Features and Neural Network Classifier, 5th International Multi-Conference on Systems, Signals and Devices,IEEE,2008

[62] M. Pechwitz and V. Margner, "HMM Based Approach for Handwritten Arabic Word Recognition Using the IFN/ENIT-Database," Proc. Int'l Conf. Document Analysis and Recognition„ 2003.

[63] S.S. Maddouri,and al,"Combination of Local and Global Vision Modeling for Arabic Handwritten Words Recognition," Proc. Int'l Conf. Frontiers in Handwriting Recognition, 2002.

[64] M. Dehghan,and al, Handwritten Farsi (Arabic) Word Recognition: A Holistic Approach Using Discrete HMM," Pattern Recognition, vol. 34, 2001

[65] A. Amin, "Recognition of Hand-Printed Characters Based on Structural Description and Inductive Logic Programming," Pattern Recognition Letters, vol. 24, 2003.

[66] M.S. Khorsheed, "Recognising Handwritten Arabic Manuscripts Using a Single Hidden Markov Model," Pattern Recognition Letters, vol. 24, 2003.

[67] S.A. Ali, Topological analysis in the design of a machine to recognise handprinted characters, Ph.D. Thesis, Brunel University, England, 1979

[68] H.Y. Abdelazim and al, "Arabic text recognition using a partial observation approach", Proc. 12th National Computer Conf., Saudi Arabia, 1990

[69] S.S. El-Dabiand al, "Arabic character recognition system: A statistical approach for recognizing cursive typewritten text", Pattern Recognition, Vol. 23, No. 5, 1990,

[70] M. Sarfraz, Offline Arabic Text Recognition system ,Proceedings of the 2003 International Conference on Geometric Modeling and Graphics (GMAG'03),IEEE,2003.

[71] G. Abandah and N. Anssari , Novel Moment Features Extraction for Recognizing Handwritten Arabic Letters,Journal of Computer Science 5 (3),ISSN 1549-3636, 2009

[72] M. Fakir and C. Sodeyama, "Machine recognition of Arabic printed scripts by dynamic programming matching method", IEICE Trans. Inform. Systems, Vol. 76, No. 2, 1993.

[73] S.A. Mahmoud, "Arabic character recognition using Fourier descriptors and character contour encoding", Pattern Recognition,1994

[74] T.S. El-Sheikh and R.M. Guindi, "Computer recognition of Arabic cursive scripts", Pattern Recognition, Vol. 21, No. 4, 1988

[75] M.F. Tolba, and al , "A recognition algorithm for printed Arabic character", Proc. IASTED lnternat. Symp. in Applied Informatics, Switzerland, 1987.

[76] A. Nouh and al, "On feature extraction and selection for Arabic character recognition", Arab GulfJ. Scient. Res., Vol. 2, No. 1, 1984.

[77] A. Nurul-Ula and A. Nouh, “Automatic recognition of Arabic characters using logic statements. Part 11. Development of recognition algorithm”, J. Engrg. Sci. King Saud Univ., Vol. 14, No. 2, 1988.

[78] F. Khella, Analysis of hexagonally sampled images with application to Arabic cursive text recognition, Ph.D. Thesis, University of Bradford, Bradford, England, 1992.

[79] A. Amin, G. Masini and J.-P. Haton, “Recognition of handwritten Arabic words and sentences”, Proc. 7th Internat. Joint Cont on Pattern Recognition, October 1984.

[80] I.S.I. Abuhaiba, Use of fuzzy set theory in pattern recognition with application to Arabic characters, Master’s Thesis, University of Bradford, England, 1990

[81] V. Margner, “SARAT - A system for the recognition of Arabic printed text”, Proc. 11th IAPR Internat. Conf on Pattern Recognition, The Hague, The Netherlands, 1992,

[82] S. Mozaffari, and al, “Structural Decomposition and Statistical Description of Farsi/Arabic Handwritten Numeric Characters,” Proc. Int’l Conf. Document Analysis and Recognition, 2005

[83] H.Y. Abdelazim and A. Abdel-Mageed, “Automatic reading of Arabic text with spell checking assistance”, Proc. Conf on the Use of Arabic Language in Information Technology, Riyadh, Saudi Arabia, May 1992 (in Arabic).

[84] H.S. Al-Yousefi and S.S. Udpa, “Recognition of handwritten Arabic characters via segmentation”, Arab GulfJ. Scient. Rex, Vol. 8, No. 2, 1990.

[85] A. Graves and J. Schmidhuber. Offline handwriting recognition with multidimensional recurrent neural networks. In 22sd Conference on Neural Information Processing Systems (NIPS), 2008.

[86] S. Alma’adeed, ”Recognition of Off-Line Handwritten Arabic Words Using Neural Network”, proc. Of the Geometric Modeling and Imaging -New Trends, 2006.

[87] S.M. Touj and N. Ben Amara, “Arabic Handwritten Words Recognition Based on a Planar Hidden Markov Model,” Int’l Arab J. Information Technology, vol. 2, 2005.

[88] A.Broumandnia and al , Handwritten Farsi/Arabic Word Recognition,IEEE, 2007.

[89] M. Kherallah and al , On-line Arabic handwriting recognition system based on visual encoding and genetic algorithm, Engineering Applications of Artificial Intelligence Elsevier, 2008.

[90] Janeway Jr, C. A. , “How the Immune System recognizes Invaders”, Scientific American, 269(3),(1993) pp. 41-47.

[91] Burnet, F. M. , “The Clonal Selection Theory of Acquired Immunity”, Cambridge University Press(1959)

[92] Tizzard, I. . Immunology: An Introduction. 2nd edition. Chap. The response of B-cells to Antigen. Pub. Saunders College.(1988a)pp. 199-223.

[93] De Castro, L. N. and Von Zuben, F. J. . Artificial Immune Systems: Part I – Basic Theory and Applications, Technical Report – RT DCA 01/99, (1999) p. 95

[94] Kepler, T and Perelson, A. . Somatic Hypermutation in B-cells : An Optimal Control Treatment. Journal of Theoretical Biology. 164.(1993) pp. 37-64.

[95] Berek, C. and Ziegner, M. . The Maturation of the Immune Response, Imm. Today, 14(8),(1993) pp. 400-402.

[96] Varela, F, and al . Cognitive Networks : Immune and Neural and Otherwise. Theoretical Immunology: Part Two, SFI Studies in the Sciences of Complexity, 2, (1988).pp.359-371.

[97] I.D. Farmer, N.H. Pachard and A.S. Perelson. " The immun system, adaptation and machine learning ". Physica D, 22 , 1986. pp 187-204

[98] Nossal, G. J. V. , "Life, Death and the Immune System", Scientific American, 269(3), (1993) pp. 21-30.

[99] Dasgupta . An overview of artificial immune systems. Artificial Immune Systems and Their Applications. pp. 3-19. Pub: Springer-Verlag.

[100] De Castro, L.N and Timmis. Artificial Immune Systems: A New Computational Intelligence Approach. Springer-Veralg. London.2002, ISBN 1-85233-594-7

[101] Timmis, J., Hone, A., Stibor, T., Clark, E., 2008. " Theoretical advances in artificial immune systems ". Volume 403, 11-32.

[102] A. Watkins, . A resource limited artificial immune classifier. MS Thesis. Mississippi State University. USA.(2001)

[103] Emma Hart,Jon Timmis,"Application areas of AIS: The past, the present and the future", Applied Soft Computing 8 (2008)pp 191–201.

[104] L. N. de Castro and F. J. von Zuben, "Learning and optimization using the clonal selection principle," IEEE Trans. Evol. Comput., vol. 6, no. 3 , Jun. 2002 , pp. 239-251.

[105] S. Forrest, S. Perelson, L. Allen, R. Cherukuri, Self–nonself-discrimination in a computer, in: Proceedings of IEEE Symposium on Research in Security and Privacy, 1994, pp. 202–212

[106] Jerne, N. K. (1974), "'Towards a Network Theory of the Immune System"', Ann. Immunol. (Inst. Pasteur) 125C, pp. 373-389.

[107] McCulloch WS, Pitts W. A logical calculus of the ideas imminent in nervous activity. Bull Math Biophys 1943;5:115-33.

[108] R.J Schalkoff, "'Artificial Neural Networks"'. McGraw-Hill New York. 1997

[109] A.K. Jain, , J. Mao, , K.M. Mohiuddin, ,. Artificial neural networks: a tutorial. Comput. IEEE March, 1996 31–44.

[110] T. Kohonen, "'Self-organization and Associative Memory,3rd Edition"', Springer, New York, 1989.

[111] J.J. Hopfield, "'Neurons with graded response have collective computational properties like those of two-state neurons"'. Proc. Natl. Acad. Sci. 81,1984, 3088–3092.

[112] Rosenblatt F, "'The perceptron: A probabilistic model for information storage and organization in the brain"'. Psychological Review 65,1958,386-408.

[113] J. Holland. " Adoption in natural and artificial systems ". The MIT press, 1975, 211.

[114] D. Golberg " Genetic Algorithm in Search, Optimisation and Machine L earning". Addison Wesley, 1989.

[115] P. W. M. Tsang, A.T.S. Au. "'A genetic algorithm for projective invariant object recognition"'. Conference proceedings, 1996 IEEE TENCON: Digital Signal Processing Applications, 1996, pp 58 - 63.

[116] Michalewiz, Z, "'Genetic Algorithms + Data Structures = Evolution Programs "', Springer Verlag, 3rd Ed,1996.

[117] Whitley."' The GENITOR algorithm and selective pressure."' In Morgan Kaufman, editor, Proceedings of the Third International Conference on Genetic Algorithms (ICGA-89), 1989 , pages 116–121.

[118] J. Greffenstette. "'Evolutionary Computation 1 : basic algorithms and operators, chapter Rank-based selection"'. Institute of Physic, 2000.

[119] J.-L. Deneubourg, S. Goss, N.R. Franks, A. Sendova-Franks,C. Detrain, et L. Chretien. "'The dynamics of collective sorting: robot-like ant and ant-like robots"'. In Proceedings of the First International Conference on Simulation of Adaptive Behavior, pp 356–365, 1990.

[120] O.Cordon, F. Herrera, and T. Stutzle, " A review on the ant colony optimization metaheuristic: Basis, models and new trends ". Mathware and Soft Computing,9(2–3),(2002), pp 141–175.

[121] A. Colorni, M. Dorigo, V. Maniezzo , "'Distributed optimisation by ant colonies"', In Proceeding of ECAL-91,Paris, France 134-142, 1991.

[122] M . Dorigo, "'Optimization, Learning and Natural Algorithms (in Italian)"'. PhD thesis, Politecnico di Milano, Italy. 1992.

[123] M. Dorigo, V. Maniezzo, A. Colorni„"' The Ant System: Optimization by a Colony of Cooperating Agents"', IEEE Transactions on Systems, Man and Cybernetics-Part B, 1996, 1(26): pp. 29-41.

[124] T. Stutzle and H. H. Hoos, "' Max–min ant system. Future Generation Computer Systems"', 16(8),2000, 889–914.

[125] M. Dorigo and L. M. Gambardella . "'Ant Colony System: A cooperating learning approach to the travelling salesman problem"'. IEEE Transactions on Evolutionary Computation, 1(1),1997, pp 53–66.

[126] B. Bullnheimer, R. F. Hartl and C. Strauss . "'A new rank-based version of the Ant System: A computational study"'. Central European Journal for Operations Research and Economics, 7(1),1996 , pp 25–38.

Annexe A

Communication ICIST'2011

Dans cette annexe nous présentons un article qui a fait l'objet d'une communication orale au "1st International Conference on Information Systems and Technologies "ICIST 2011"" 24, 25 et 26 avril 2011, Tébessa (Algérie). Dans cette article nous avons présenté l'essentiel de notre contribution.

Article:

Meriem GAGAOUA, Hamza GHILAS Abdelouahab MOUSSAOUI, *"Approche hybride bio-inspirée pour la reconnaissance de l'écriture arabe manuscrite"*, "ISICT'2011", 24, 25 et 26 avril 2011, Tébessa (Algérie)

Approche hybride bioinspirée pour la reconnaissance de l'écriture arabe manuscrite

GAGAOUA Meriem[(1)], GHILAS Hamza[(2)], MOUSSAOUI Abdelouahab[(3)]
Département informatique
Université Ferhat Abas, Sétif
Sétif, Algérie
gagaouameriem@yahoo.fr [(1)],hamzaghilas@yahoo.fr [(2)] , mousaoui.abdel@gmail.com [(3)]

***Abstract*— Dans cet article nous présentons une approche hybride bioinspirée qui se base sur les systèmes immunitaires artificiels et les algorithmes génétiques pour la reconnaissance de l'écriture arabe manuscrite. Nous avons utilisé un système immunitaire artificiel pour la reconnaissance des mots, puis on a appliqué un optimiseur génétique pour sélectionner la combinaison d'attributs appropriés pour la classification immunitaire. L'ensemble des testes s'est effectué sur la base de données des villes tunisienne IFN/ENIT.**

Mots clé : Reconnaissance de l'écriture arabe manucrite, systemes immunitaires artificiels, algorithmes genetiques.

I. INTRODUCTION

Toute information écrite peut être reprise et informatisée à différentes fins et dans plusieurs domaines (la bureautique, la reconnaissance de montants littéraux des chèques bancaires, le tri automatique du courrier et etc.). Pour cela la reconnaissance optique des caractères OCR (*Optical Character Recognition*) qui dérive du domaine de la reconnaissance de forme occupe une place importante dans la recherche scientifique.

Le manuscrit de la langue arabe est particulier et différent de celui des langues latines, cependant la conception d'un OCR pour cette langue est une véritable problématique. Plusieurs recherches se sont dirigées vers cet axe et plusieurs techniques de classification de données sont appliquées à la reconnaissance de l'écriture arabe: les réseaux de neurones artificiels [2,3,4], les modèles de Markoves cachés(MMC)[5], les algorithmes génétiques [6], la logique flou [1] et etc.

Ces méthodes ont leurs avantages mais révèlent une insuffisance envers la nature complexe de l'écriture arabe (cursive, présence des points diacritique, pseudo mots ligaturés et etc.).

Les systèmes immunitaires naturels sont très puissants par rapport à leurs responsabilités dans le fonctionnement des corps biologiques, ils ont les particularités d'adaptabilité, d'auto-organisation, de mémorisation, de reconnaissance et d'évolution dans leurs environnements complexes. Le pouvoir de reconnaissance des ses systèmes représente une bonne source d'inspiration pour la conception d'un classifieur pour les données complexes tel que le manuscrit de la langue arabe.

Nous présentons dans cet article une approche hybride basée sur les algorithmes bio-inspirés. Nous avons couplé les systèmes immunitaires artificiels (AIS) avec les optimiseurs génétiques pour tirer parti de leurs avantages dans le but d'améliorer le taux de reconnaissance de l'écriture arabe manuscrite.

La reconnaissance des mots est assurée par un système immunitaire artificiel, mais le choix des attributs pour la classification n'est pas évident. Pour la sélection d'une combinaison d'attributs qui donne le meilleur taux de reconnaissance nous avons utilisé les algorithmes génétiques qui sont réputé d'être des bons optimiseurs.

Dans la première section nous présentons l'écriture manuscrite arabe, dans la deuxième les systèmes immunitaire et les algorithmes génétiques, dans la troisième nous présentons notre algorithme bioinspiré et dans la dernière section nous présentons les résultats de la classification.

II. RECONNAISSANCE DE L'ECRITURE ARABE

La reconnaissance des textes cursifs reste toujours un problème ouvert aussi bien dans sa forme manuscrite que imprimée. Ceci à cause des difficultés auxquelles sont confrontés les chercheurs et les développeurs, telles que la variabilité de la forme, du style, et de l'inclinaison de l'écriture arabe manuscrite.

Contrairement au latin, la reconnaissance de l'écriture arabe imprimée ou manuscrite reste encore aujourd'hui un domaine non complètement exploré. A. nazif [8] fus le premier à travailler sur la reconnaissance de l'écriture arabe dans sa thèse de master en 1975.

Les OCRs différent suivant le mode d'acquisition des données (" en-ligne " ou " hors-ligne "), le mode d'écriture traitée (imprimée ou manuscrite), selon que l'analyse s'opère sur la totalité du mot ou par segment composé d'un seul caractère (les approches globales ou analytiques).

A. Description et caractéristiques de l'ecriture arabe

La langue arabe est une langue universelle, c'est la langue officielle de 25 payes de plus de 250 millions de population. Ses caractères sont utilisés dans d'autres langues telles que farsi, jawi, et etc. Les musulmans peuvent lire et écrire la langue arabe car c'est la langue du Coran, le saint livre des musulmans.

Les caractères de l'alphabet arabe ont différentes formes dans le mot (isolé, début, milieu, fin). Exemple : مـ/ ـمـ / ـمـ/م
Les caractéristiques de la langue arabe peuvent être résumées dans les points suivants[9] :

- L'arabe s'écrit de droite à gauche.
- Un mot peut être constitué de deux ou de plusieurs sous mot et suit une ligne imaginaire appelée ligne de base.
- Quinze caractères ont des points qui peuvent être au dessous ou au dessus de la partie principale du mot.
- L'arabe est toujours écrite cursivement et les mots sont séparés par des espaces sauf avec les caractères (ز،ا،ر،و،د،ذ).

B. *Le processus de reconnaissance de l'écriture arabe*

Le processus de reconnaissance de l''ecriture arabe suit les phases suivantes :

1) Prétraitement

C'est l'ensemble des opérations qui apportent des transformations à une image pour réduire les bruits superposés aux données. les principales opérations sont la binarisation [10], la squelettisation [11], le redressement de l''ecriture [12], la normalisation et la d´détection de la ligne de base[13].

2) Segmentation

La segmentation vise à diviser le texte en unités simples. Elle peut être implicite où les caractères sont segmentés lors de la reconnaissance [14] ou explicite réalisé avant la reconnaissance [2][7]. Cette phase est très complexe et engendre des erreurs à cause de la difficultés de trouver les vrais limites des caractères [15]. Quelques systèmes de reconnaissance procèdent sans segmentation des mots suivant l'approche holistique ou le mot est reconnu entièrement en passant directement à la phase d'extraction des caractéristiques [3][5].

3) Extraction des caractéristiques

C'est l'une des étapes les plus délicates et les plus importantes. La reconnaissance d'un caractère passe d'abord par l'analyse de sa forme et l'extraction de ses traits caractéristiques (primitives) qui seront exploités pour son identification. Les caractéristiques peuvent être classées en quatre groupes principaux : des caractéristiques structurelles qui d écrivent la topologie d'une forme et sa géométrie en donnant ses propriétés globales et locales tel que les boucles, les voyellations et les zigzags (hamza,٩), le nombre de points diacritiques et leurs positions et etc[4][5], des caractéristiques statistiques qui décrivent une forme en terme d'un ensemble de mesures extraites à partir de cette forme, on trouve le zonage[16] la méthode de lieu géométrique (Loci)[17] et les moments[2] et des caractéristiques basées sur les transformations globales [3] et superposition des modèles et corrélation [8].

4) La classification et la reconnaissance

C'est la phase de prise de décision, elle utilise les caractéristiques extraites dans la phase précédente pour la reconnaissance du textes. Il existe plusieurs approches de classification, l'approche structurelle qui repose sur la structure physique du caractère tel que les arbres de décision et les règles de production, l'approche statistique tel que les KNNs, les classifieurs bayésien et d'autres approches qui se basent sur les chaines de markove, les réseaux de neurones et les algorithmes génétique.

Le tableau suivant présente les performances et les caractéristiques de quelques systèmes de reconnaissance de l'écriture arabe.

TABLE I. PERFORMANCES ET CARACTERISTIQUES DE QUELQUES SYSTEMS DE RECONNAISSANCE DE L'ECRITURE ARABE

Référence	Type de classifieur	Type d'écriture	Segmentation	Caractéristiques	Taux de reconnaissance
[2]	Réseaux de neurones	Texte imprimé, hors ligne	Oui	Statistiques (moments de HU)	73%
[3]	Réseaux de neurones	Texte manuscrit hors ligne	Non	Transformation globale (transformée en ondelette)	95,8%
[4]	Réseaux de neurones	Texte manuscrit hors ligne	Oui	Structurelles (points finale / branchement/ croisement, boucle, longueur, largeur)	73%
[6]	Algorithmes génétique	Texte manuscrit en ligne	Non	Chaîne de code	97%
[5]	Chaine de Markov	Texte manuscrit hors ligne	Non	Structurelles (Boucles, points diacritiques, points finale/branchement/croisement)	87%
[7]	Classifieur bayesien	Texte imprimé hors ligne	Oui	Statistiques (Caractéristique Loci)	98%

III. ALGORITMES BIOINSPIRÉS

Pendant les derrières décennies l'inspiration des systèmes informatiques de la biologie a prouvé sont succès dans la résolution des problèmes complexes ce qui a donné naissance aux algorithmes bioinspirés tel que les systèmes immunitaires artificiels, les algorithmes génétiques, les réseaux de neurones, les fourmis artificiels et etc.

A. Les systémes imuunitaires artificiels

Un système immunitaire naturel (SIN) est un système de défense, il protège le corps des vertébrés contre des agents infectieux (antigène) tel que les virus, les bactéries et d'autres parasites. Le SIN se met à contribution, lorsque les cellules qui le composent (anticorps) détectent la présence d'un antigène.

Les systèmes immunitaires artificiels AIS (*Artificial immun System*) sont des systèmes de classification et de reconnaissance adaptatifs, inspirés des fonctions, des principes et des modèles de l'immunologie théorique, qui sont appliqués à la résolution des problèmes [18]. Leurs origines reviennent aux travaux théoriques d'immunologie [19,20].

Les AISs visent à développer un modèle informatique qui préserve les caractéristiques importantes des SINs, ils apportent, d'un point de vue informatique, des propriétés telles que la reconnaissance, la discrimination, la mémorisation, l'apprentissage, l'auto-organisation, l'adaptation, la robustesse et l'évolutivité. En générale il ya trois composantes principales dans l'ingénierie immunitaire [21]:

- Un modèle pour la représentation de la problématique.
- Un mécanisme pour l'évaluation de l'affinité entre les anticorps et les antigènes.
- Un algorithme pour le contrôle du système (sélection clonal, sélection négative, réseaux immunitaire).

A l'heure actuelle, on trouve l'utilisation des systèmes immunitaires artificiels dans plusieurs domaines : détection d'anomalies, détection d'intrusions, reconnaissance de formes, classification et etc.

Par rapport à d'autres algorithmes bioinspirés, les AIS offrent des modèles et des algorithmes différents (la sélection clonale, sélection négative, réseaux immunitaires) conçus pour différentes applications.

Dans notre approche on s'intéresse à la sélection clonale qui est bien utilisée pour la reconnaissance de forme duquel dérive notre application.

1) Algorithme de la selection clonale

Cet algorithme utilise la propriété de la sélection clonale réalisée par les cellules immunitaires et la maturation d'affinité. [22]

Début
  P = un ensemble de forme à reconnaitre
  M = population d'individus aléatoires
  tant que (*une forme minimale n'est pas reconnue*) **faire**
    Pour (*chaque forme* P_i) **faire**
      aff= affinité(P_i, M_i)
    fin
    Sélectionner n_1 éléments ayant la meilleure affinité avec les éléments de M
    Générer des copies de ces éléments proportionnellement à leur affinités avec l'antigène
    Muter toutes les copies proportionnellement avec leur affinité avec les formes de l'ensemble P(plus l'affinité est élevée, plus la mutation est faible)
    Ajouter les individus muté dans la population M
    Choisir n_2 de ces éléments mutés (optimisés) comme mémoire
  fin
fin

Figure 1 Algorithme de la sélection clonal

B. Les algorithmes genetiques

Les algorithmes génétiques (AG) sont des algorithmes d'exploration fondés sur les mécanismes de la reproduction naturelle et de la génétique. Ils se basent sur la théorie de Darwin « survie des individus les mieux adaptés ». Ces algorithmes sont initialement introduits par John Holland [23] qui a développé leurs principes fondamentaux, puis Goldberg [24] les a utilisés pour résoudre des problèmes d'optimisation.

Les AGs ont prouvé leur succès dans les problèmes d'optimisation à large espace de solutions [25]. Ils sont utilisés lorsque la recherche exhaustive d'une solution est couteuse en termes de temps d'exécution.

1) Caractéristique des AG

Les AG ont les caractéristiques suivantes :

- Ils utilisent un codage des paramètres, et non les paramètres eux-mêmes.
- Ils travaillent sur une population d'individus, au lieu d'un individu unique.
- Ils utilisent une fonction d'évaluation.
- Ils utilisent des règles de transition probabilistes et non déterministes.

2) Principe de fonctionnement des AG

L'AG présenté par l'organigramme de la figure 2, débute par la génération d'une population initiale de N individus, pours lesquels les valeurs de leurs fonctions objectives seront calculées, les individus seront sélectionnés par une méthode de sélection (roulette, aléatoire, rang ….). Les individus concernés par l'opérateur de croisement seront choisis selon une probabilité Px. Leurs résultats peuvent être mutés par un opérateur de mutation avec une probabilité de mutation Pm. Les individus issus de ces opérateurs génétiques seront insérés par une méthode d'insertion dans la nouvelle population dont la valeur de la fonction objective de chaque individu sera évaluée. Un test d'arrêt sera effectué pour vérifier la qualité des individus obtenus .Si ce test est vérifié alors l'algorithme s'arrête avec une solution optimale, sinon le processus sera réitérer pour la nouvelle population.

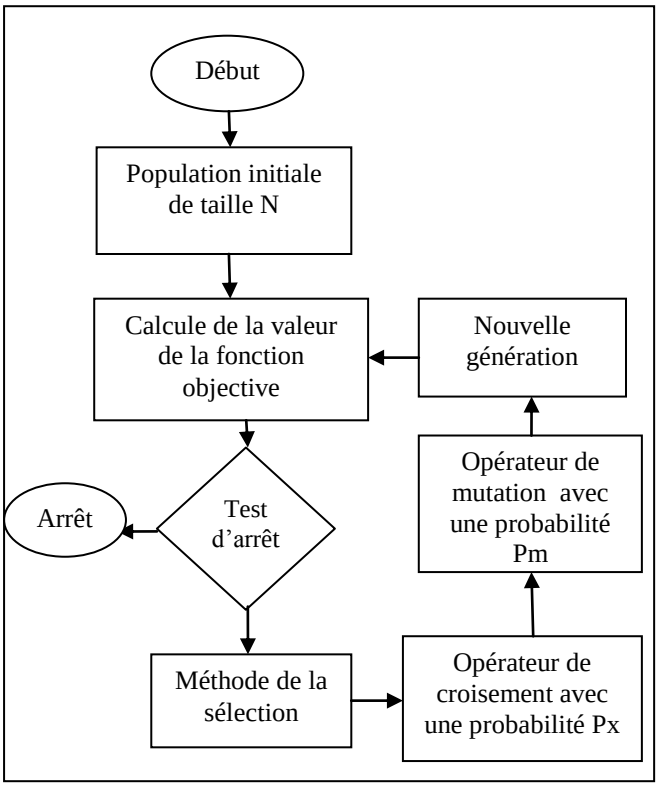


Figure 2 Organigramme de l'algorithme génétique [25]

IV. ALGORITHEME HYBRIDE IMINO-GENETIC

Plusieurs travaux traitant la reconnaissance de l'écriture arabe ont été réalisés mais le taux de reconnaissance reste toujours insatisfaisant. Ce dernier dépend du classifieur et des attributs de classification utilisés.

Pour avoir un taux de reconnaissance élevé l'idée est de concevoir un algorithme hybride bioinspiré pour tirer avantages de leurs efficacités.

Comme classifieur on a choisi d'utiliser les systèmes immunitaires artificiels qui ont la faculté de mémorisation et de bon apprentissage, de plus ils sont bien utilisés dans le domaine de reconnaissance de formes.

Dans la littérature il existe plusieurs types d'attributs pour la classification. Dans notre approche nous utilisons les caractéristiques structurelles d'un mot, qui sont nombreuses et la recherche exhaustive pour trouver la bonne combinaison de caractéristiques est très couteuse en temps de calcule. Pour remédier à cette lacune, notre choix c'est porté sur les AG qui s'appliquent souvent pour les problèmes d'optimisation.

Nous proposons d'utiliser un maximum d'attributs structurels pour la classification, par exemple :

J	H	PPD	NPD	B	NSM	NCS	NC		

J : Jambe, H : Hampe ; PPD : préséance de points diacritique, NPD : nombre de points diacritique, B : boucle, NSM : nombre de sous mots, NCS : nombre de caractères dans chaque sous mot, NC : nombre de caractères,

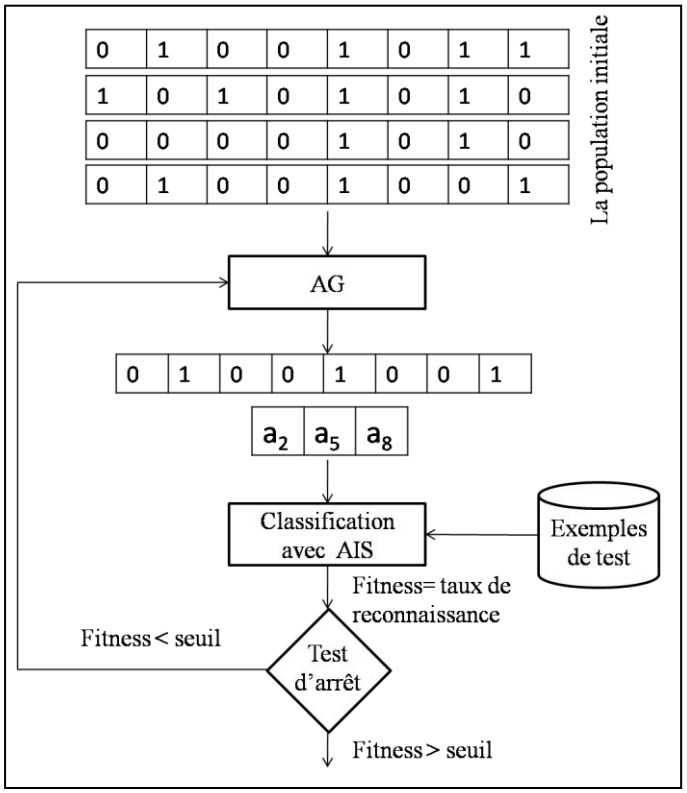


Figure 3 Algorithme imino-genetic

La figure 3 est une illustartion du fonctinnement générale de notre algorithme.

L'AG reçoit en entrée un ensemble de vecteurs d'attributs avec un codage binaire, 1 signifie la prise en compte de l'attribut corespendant et 0 pour le cas contraire. A chaque iteration l'AG selectionne un vecteur V de la population, puis l'algorithme de classification immunitaire sera lancé avec les attributs corespendants au vecteur slectionné precedement et le taux de reconnaissance pour cet ensemble d'attributs sera la valeur de la fonction fitness pour ce vecteur V.

L'opération est réitérée jusqu'à ce que le critère d'arrêt de l'AG soit atteint (nombre d'itération). A la fin de l'algorithme on aura en sortie le meilleur taux de reconnaissance qui correspond à la combinaison d'attributs optimale.

A. Pour l'AG

- La population initiale est constituée d'un ensemble de vecteurs représentants des combinaisons d'attributs.
- Le codage des vecteurs est binaire.
- La fonction fitness est calculée à partir du taux de classification pour un vecteur d'attributs donné.

B. Le classifieur immunitaire

- Les antigènes sont les mots à reconnaitre, ils sont représentés par un vecteur de caractéristiques, ces dernières sont choisies par l'algorithme génétique.
- Les anticorps et les cellules mémoire sont représentés par des vecteurs de même type que ceux des antigènes.
- L'affinité entre les anticorps et les antigènes est calculée avec une distance correspondante aux données traitées.
- L'antigène est affilié à la classe avec laquelle il aura la valeur d'affinité la plus petite.

V. IMPLÉMENTATION ET ÉVALUATION DES RÉSULTATS

A. L'environnement de dévellepemnt

Pour le teste de notre algorithme imino-genetic, on a utilisé l'environnement Windows, et le langage de programmation java avec l'IDE eclipse. L'implémentation du classifieur immunitaire est réalisé avec l'API weka qui offre un package très riche pour les systèmes immunitaires artificiels *weka.classifiers.immune*. le Framework JGAP(java genetic algorithm programming) et utilisé pour l'optimisation génétique .

B. La base d'image :

L'ensemble des testes s'est porté sur un sous ensemble de la base des mots manuscrits des villes tunisiennes IFN /ENIT. Cette base est développée par l'institut des technologies de communications (IfN) en coopération avec l'école nationale d'ingénieurs de Tunis (ENIT) en 2002, elle contient l'ensemble des noms des villes tunisiennes.

C. Extraction des attributs

Nous avons proposé une liste de 15 attributs présentés dans le tableau ci-après, ces attributs sont calculés à partir des images squelettisées.

PJ	**PH**	**PB**	**PPD**	**NbCV**	**NbJ**	**NbH**	**NbPDU**	

PPDD	**PPDT**	**NPDT**	**NBU**	**NBD**	**NBSM**	**NBC**

PJ : Présence de Jambes, **PH** : Présence de hampe, **PB** : Présence de boucle, **NbJ** :nombre de Jambes, **PH :** présence de Hampe , **NbH** nombre de Hampes , **PPD** :présence de points diacritique, **NbPDU :** nombre de points diacritique unique, **PPDD :** nombre de points diacritique double, **PPDT** :nombre de points diacritique triple, **NBU** :nombre de boucle unique, **NBD :** nombre de boucle double, **NBSM :**nombre de sous mots, **NBC** :nombre de caractères, **PC** : présence de concavités, **NBCV** : nombre de concavités.

D. Classification et opptimisation

L'algorithme imino-genetic utilisé se base sur le système immunitaire et les algorithmes génétiques.

Pour AG :

- Chaque individu (chromosome) de la population représente une solution candidate pour la sélection de sous ensemble de caractéristiques candidates, 2^{15} combinaisons possibles.
- un chromosome contient 15 genes, un gene pour chaque caractéristique qui peut prendre deux valeurs 0 / 1.
- La population initiale est générée aléatoirement, un point de croisement et de mutation binaire est utilisé, la méthode de sélection utilisé est celle de la roulette, la taille de la population initiale 20, le nombre de générations 1000
- La probabilité de croisement est 0.8, la probabilité de mutation est 0.2. La fonction d'évaluation (fitness) est calculée à partir du taux de classification pour un vecteur d'attributs donné.

Pour l'immun :

- Les antigènes et les anticorps sont codés par des entiers,
- L'affinité entre les anticorps et les antigènes est calculée avec la distance euclidienne.
- L'algorithme immunitaire utilisé est CLONALG.

E. Résultats et discussions

L'utilisation de la totalité des attributs avec l'algorithme de classification CLONALG a généré un taux de

reconnaissance de 68%. En appliquant l'algorithme hybride le taux de reconnaissance est amélioré, l'algorithme génétique a sélectionné le vecteur présenté ci-dessus qui donne le meilleur taux de reconnaissance de 70%.

0	0	0	1	0	1	1	1	0	0	1	1	1	1	1

Les taux de reconnaissance générés ne sont pas satisfaisant, ceci est dû à la forte cursivité et à la variabilité de l'écriture manuscrite arabe d'un scripteur à un autre tel que la boucle qui n'apparaît pas dans l'image suivante :

VI. CONCLUSION

Dans cet article nous avons proposé un algorithme hybride bioinspiré, basé sur les systèmes immunitaires comme outil de classification pour la reconnaissance de l'écriture arabe manuscrite et les AG comme optimiseur appliqué sur l'ensemble des attributs caractérisant les mots à reconnaitre. Nous avons appliqué notre algorithme pour la reconnaissance des mots à savoir l'ensemble des villes tunisiennes. Cette application peut présenter beaucoup d'intérêts dans le tri automatique du courrier et la reconnaissance des montants des chèques. Pour généraliser notre contribution pour la reconnaissance d'un texte arabe manuscrit quelconque nous envisageons d'introduire l'aspect sémantique à travers un dictionnaire ou une ontologie.

REFERENCES

[1] I.S.I. Abuhaiba,'' Use of fuzzy set theory in pattern recognition with application to Arabic characters'', Master's Thesis, University of Bradford, England, 1990.

[2] Fahmy, S.Al Ali "Automatic recognition of handwritten Arabic characters using their geometrical features" . Studies in informatics and control journal (SIC journal), vol.10, No 2, 2001

[3] A.Broumandnia and al , "Handwritten Farsi/Arabic Word Recognition",IEEE, 2007.

[4] R. Haraty and C. Ghaddar, "Arabic Text Recognition," Int'l Arab J.Information Technology, vol. 1,No 2, 2004.pp 156,163.

[5] M.S. Khorsheed, "Recognising Handwritten Arabic Manuscripts Using a Single Hidden Markov Model,"elseiver Pattern Recognition Letters, vol. 24, 2003.

[6] M. Kherallah and al , "On-line Arabic handwriting recognition system based on visuencoding and genetic algorithm", Engineering Applications of Artificial Intelligen, Elsevier, 2008.

[7] H.Y. Abdelazim and al, "Arabic text recognition using a partial observation approach", Proc. 12th National Computer Conf., Saudi Arabia, Saudi Arabia, 21-24 October 1990, pp. 427437.

[8] A. Nazif, "A system for the recognition of the printed Arabic characters", Master's Thesis, Faculty of Engineering, Cairo University, 1975.

[9] A. Ali Aburas et M. E. Gumah,"'Arabic Handwriting Recognition: Challenges and Solutions"', .IEEE , 2008.

[10] B. Al-Badr, S.A. Mahmoud :" Survey and bibliography of Arabic optical text recognition" . Signal processing, vol. 41,elsevier, 1995.

[11] A. Nouh and al , "A proposed algorithm for thinning binary Arabic character patterns", Proc. 1st Kuwaiti Computer Conf, Kuwait, March 1989.

[12] T. Steinherz, E. Rivlin, N. Intrator: "Off-line cursive word recognition: a survey ". International journal on document analysis and recognition, 1999.

[13] A. AL-Shatnawi and K. Omar," A Comparative Study between Methods of Arabic Baseline Detection", International Conference on Electrical Engineering and Informatics Selangor, Malaysia,IEEE, 2009.

[14] A. Cheung, M. Bennamoun and N. W. Bergmann, "'An Arabic Optical Character Recognition System using Recognition-Based Segmentation"', Pattern Recognition,2001. pp.215-233,

[15] A. Belaid, "Analyse et reconnaissance de documents, Cours INRIA: le Traitement électronique de Documents", Collection ADBS, 3-7 octobre, Aix-en-Provence, 2002.

[16] S.A. Ali, "Topological analysis in the design of a machine to recognise handprinted character", Ph.D. Thesis, Brunel University, England, 1979.

[17] H.Y. Abdelazim and al, "Arabic text recognition using a partial observation ap proach", Proc. 12th National Computer Conf., Saudi Arabia, 1990 .pp. 427437.

[18]] L.N. de Castro and J. Timmis. Artificial Immune Systems: a new computational intelligence approach. Springer, 2002.

[19]] I.D. Farmer, N.H. Pachard and A.S. Perelson. « The immun system, adaptation and machine learning ». Physica D, 22 , 1986.pp 187-204

[20]] F. Varela, A. Coutinho, B .Dupire, and N. Vaz. « cognitive Network : Immun, Neural and otherwise ». Théoretical Immunology, ,1988., pp 359-375

[21] Timmis, J., Hone, A., Stibor, T., Clark, E., 2008. « Theoretical advances in artificial immune systems ». Volume 403, 11–32.

[22] L. N. de Castro and F. J. von Zuben, "Learning and optimization using the clonal selection principle," IEEE Trans. Evol. Comput., vol. 6, no. 3 pp. 239–251, Jun. 2002.

[23] J. Holland. « Adoption in natural and artificial systems ». The MIT press, 1975, 211.

[24] D. Golberg " Genetic Algorithm in Search, Optimisation and Machine L earning". Addison Wesley, 1989.

[25]] P. W. M. Tsang, A.T.S. Au. A genetic algorithm for projective invariant object recognition. Conference proceedings, 1996 IEEE TENCON: Digital Signal Processing Applications, 1996, pp 58 – 63.

Annexe B

Communication JDLIO'2011

Dans cette annexe nous présentons un autre article qui a fait l'objet d'une communication orale au journées doctorales de oran "JDLIO 2011".

Article:

Meriem GAGAOUA, Abdelouahab MOUSSAOUI, *"Reconnaissance de L'Ecriture Arabe Manuscrite par un Classifieur Neuro-Génétique "*, 1^{ere} journées doctorales du laboratoire d'informatique d'Oran, "JDLIO'2011", les 31 Mai et 01 Juin, Oran-Algérie, 2011.

Reconnaissance de L'Ecriture Arabe Manuscrite par un Classifieur Neuro-Génétique

Meriem GAGAOUA [(1),] Hamza GHILAS [(2),] Abdelouahab MOUSSAOUI [(3)]
Département informatique
Université Ferhat Abas, Sétif
Algérie
gagaouameriem@yahoo.fr [(1)]hamzaghilas@yahoo.fr [(2)] mousaoui.abdel@gmail.com[(3)]

Résumé. Dans cet article nous présentons notre contribution à la reconnaissance de l'écriture manuscrite arabe en utilisant des algorithmes bioinspirés. Elle est basée sur les réseaux de neurones (RN) pour la reconnaissance des mots arabe et les algorithmes génétiques pour la sélection des attributs de classification approprié au RN et donnant le meilleur taux de reconnaissance. L'ensemble des testes s'est effectué sur la base de données des villes tunisienne IFN/ENIT. Le taux de reconnaissance généré est satisfaisant par rapport à la complexité du manuscrit arabe ainsi que la liste des attributs pertinents est retenue.

Mots clés: Reconnaissance de l'écriture rabe manuscrite, réseaux de neurones , algorithmes génétique.

1 Introduction

La reconnaissance optique des caractères OCR (Optical Character Recognition) dérive du domaine de la reconnaissance de forme et occupe une place importante dans la recherche scientifique. La conception d'un OCR pour le manuscrit arabe est une véritable problématique. Plusieurs techniques de classification de données sont appliquées à la reconnaissance de l'écriture arabe (REA) : les réseaux de neurones artificiels [1,2], les modèles de Markoves cachés(MMC) [3], les algorithmes génétiques [4], et etc. Ces méthodes ont leurs avantages mais révèlent une insuffisance envers la nature complexe de l'écriture arabe (cursive, présence des points diacritique, pseudo mots ligaturés et etc.).

Dans le but d'améliorer le taux de reconnaissance de l'écriture arabe manuscrite, nous présentons dans cet article une approche basée sur les algorithmes génétique et les réseaux de neurones. La reconnaissance des mots est assurée par un réseau de neurones, mais le choix des attributs pour la classification n'est pas évident. Pour la sélection d'une combinaison d'attributs pertinents et générant un taux de reconnaissance élevé et satisfaisant, nous avons utilisé les algorithmes génétiques qui sont réputés d'être des bons optimiseurs et bien appliqué dans ce genre de problèmes.

Dans la première section nous présentons l'écriture manuscrite arabe, dans la deuxième les réseaux de neurones et les algorithmes génétiques, dans la troisième nous présentons notre algorithme Neuro-Génétique et dans la dernière section nous présentons les résultats.

2 Etat de l'art

Contrairement au latin, la reconnaissance de l'écriture arabe imprimée ou manuscrite reste encore aujourd'hui un domaine non complètement exploré. A. nazif [6] fus le premier à travailler sur la reconnaissance de l'écriture arabe dans sa thèse de master en 1975.

2.1 Caractéristiques de l'écriture arabe

La langue arabe est une langue universelle, c'est la langue officielle de 25 payes. Les musulmans peuvent lire et écrire la langue arabe car c'est la langue du Coran, le saint livre des musulmans.

Les caractéristiques de la langue arabe peuvent être résumées dans les points suivants [5] :

- L'arabe s'écrit de droite à gauche.
- Un mot peut être constitué de deux ou de plusieurs sous mots et suit une ligne imaginaire appelée ligne de base.
- Quinze caractères ont des points qui peuvent être au dessous ou au dessus de la partie principale du mot.
- L'arabe est toujours écrit cursivement et les mots sont séparés par des espaces sauf avec les caractères (ز،ا،ر،و،د،ذ).

2.2 Le processus de reconnaissance de l'écriture arabe

Le processus de reconnaissance de l'écriture arabe suit les phases suivantes : Prétraitement, segmentation, extraction de caractéristiques et la classification ou la reconnaissance.

Le tableau suivant présente les performances et les caractéristiques de quelques systèmes de reconnaissance de l'écriture arabe.

Tableau 1. Performances et caractéristiques de quelques Systems de reconnaissance de l'écriture arabe

Réf	Classifieur	Type d'écriture	Caractéristiques	Taux de reconnaissance
[1]	Réseaux de neurones	Texte manuscrit hors ligne	Transformation globale (transformée en ondelette)	95,8%
[2]	Réseaux de neurones	Texte manuscrit hors ligne	Structurelles (points finale / branchement/ croisement, boucle, longueur, largeur)	73%

[4]	Algorithmes génétique	Texte manuscrit en ligne	Chaîne de code	97%
[3]	Chaine de Markov	Texte manuscrit hors ligne	Structurelles (Boucles, points diacritiques, points…)	87%
[5]	Classifieur bayesien	Texte imprimé hors ligne	Statistiques (Caractéristique Loci)	98%

3 Les réseaux de neurones et les algorithmes génétiques

Pendant les derrières décennies l'inspiration des systèmes informatiques de la biologie a prouvé son succès dans la résolution des problèmes complexes, ce qui a donné naissance aux algorithmes bioinspirés tel que les réseaux de neurones, les systèmes immunitaires artificiels, les algorithmes génétiques, les essaimes de particules etc.

3.1 Les réseaux de neurones

Parmi les algorithmes s'inspirant de la biologie, on trouve les Réseaux de neurones(RN) qui représente une tentative de reproduire artificiellement le fonctionnement du cerveau humain. En 1943, Warren McCulloch et Walter Pitts en s'inspirant de leurs travaux sur le neurone biologique, ont proposé un des premiers modèles de neurone artificiel [8] qui deviendra la base des réseaux de neurones artificiels. Il existe plusieurs modèles de RN reflétant différentes topologies. Dans notre approche nous nous intéressons au perceptron multicouche.

3.1.1 Le perceptron multicouche

Le perceptron multicouche [9] (MLP multilayer perceptron) est un réseau orienté de neurones artificiels organisé en couches où l'information circule dans un seul sens, de la couche d'entrée vers la couche de sortie. Un perceptron multicouche contient généralement trois couches, une couche d'entrée, une couche de sortie et une couche intermédiaire appelée couche cachée. Les neurones de la couche d'entrée sont passifs, ils n'altèrent pas les informations, ils reçoivent les valeurs des attributs et les transmettent à la couche cachée. Cette dernière est la couche la plus importante dans le MLP, ses neurones sont connectés aux neurones des autres couches, c'est la couche responsable de l'apprentissage du réseau. Ce modèle de réseau de neurones est très utilisé pour la classification de données. L'architecture du réseau dépend de la nature du problème traité, le nombre d'attributs de classification détermine le nombre de neurones de la couche d'entrée et la couche de sortie aura autant de neurones que de classes à attribuer aux individus.

3.2 Les algorithmes génétiques

Les algorithmes génétiques (AG) sont des algorithmes d'exploration fondés sur les mécanismes de la reproduction naturelle et de la génétique. Ces algorithmes sont initialement introduits par John Holland [10] qui a développé leurs principes fondamentaux, puis Goldberg [11] les a utilisés pour résoudre des problèmes d'optimisation.

Le principe générale de fonctionnent des AG est présenté dans la figure Fig.1

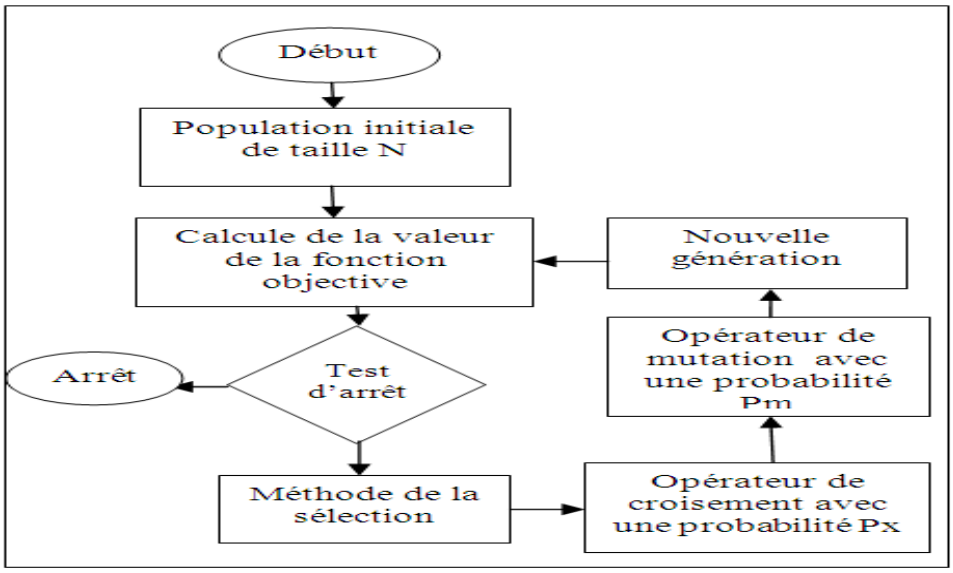


Fig.1 Organigramme de l'algorithme génétique.

4 Contribution

Notre contribution pour la reconnaissance de l'écriture manuscrite arabe se focalise sur l'algorithme de classification neuronale MLP et l'algorithme génétique utilisé pour la sélection d'attributs de classification qui génèrent le meilleur taux de reconnaissance des mots.

Nous proposons d'utiliser la liste suivante de caractéristiques structurelles des mots comme attributs de classification :

PJ	NbJ	PH	NbH	PB	NbBU	NbBD	PPD	

PPDD	PPDT	NbPDT	NbPDU	NbPDD	NBSM	NBC	NbCV

PJ : Présence de Jambes, **PH** : Présence de hampe, **PB** : Présence de boucle, **NbJ** :nombre de Jambes, **NbH** : nombre de Hampes , **PPD** :présence de points diacritique, **NbPDU** : nombre de points diacritique unique**, NbPDD** : nombre de points diacritique doubles **PPDD** : présence de points diacritique double, **PPDT** :présence de points diacritique triple, **NbBU** :nombre de boucle unique, **NbBD** : nombre de boucle double, **NBSM** :nombre de sous mots, **NBC** :nombre de caractères, **PC** : présence de concavités, **NBCV** : nombre de concavités**.**

Le principe du fonctionnement général de notre algorithme Neuro-Génétique est présenté dans la figure Fig.2. Initialement, l'AG reçoit en entrée un ensemble de vecteurs d'attributs avec un codage binaire, 1 signifie la prise en compte de l'attribut correspondant et 0 pour le cas contraire. A chaque itération l'AG sélectionne un vecteur V de la population, puis l'algorithme de classification neuronale sera lancé avec les attributs correspondants au vecteur sélectionné précédemment, le taux de reconnaissance pour cet ensemble d'attributs sera la valeur de la fonction fitness pour ce vecteur V dans l'AG.

Ce procédé est réitéré jusqu'à ce que le critère d'arrêt de l'AG soit atteint (fitness supérieur à un seuil donné). A la fin de l'algorithme on aura en sortie le meilleur taux de reconnaissance qui correspond à la combinaison d'attributs optimale.

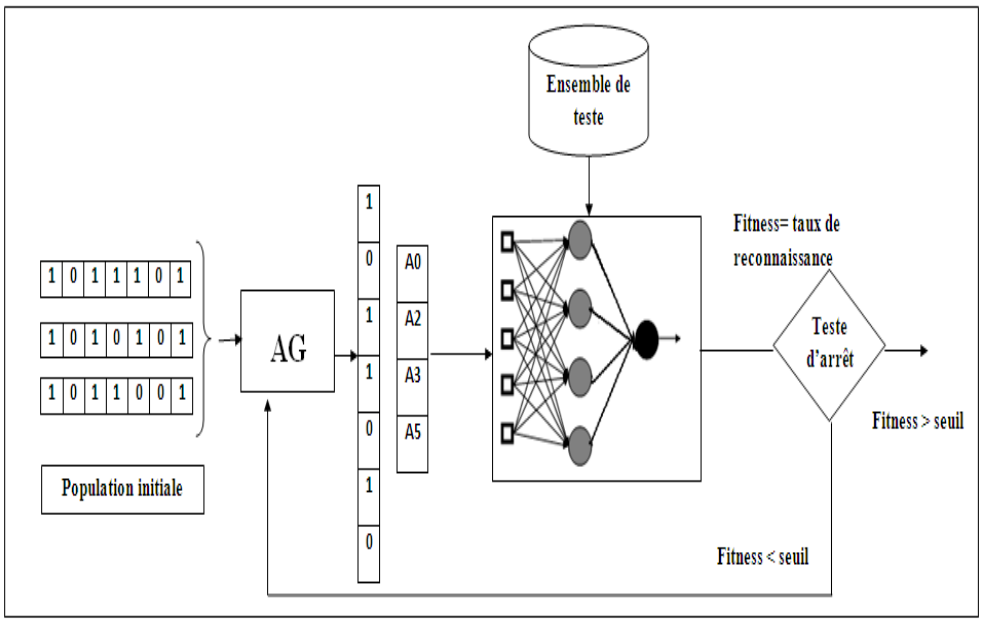


Fig.2 **Organigramme de l'algorithme Neuro-Génétique**.

5 Résultats et discussion

Pour le teste de notre algorithme, nous avons utilisé l'environnement Windows, et le langage de programmation java. L'algorithme MLP de l'API weka et le Framework JGAP (java genetic algorithm programming) ainsi que la base d'image IFN/ENIT.

L'utilisation de la totalité des attributs de classification proposés avec l'algorithme de classification MLP a généré un taux de reconnaissance de 68,2%. En appliquant notre algorithme hybride le taux de reconnaissance est amélioré, l'algorithme génétique a sélectionné le vecteur présenté ci-dessus qui donne le meilleur taux de reconnaissance de 70 ,84%.

0	1	0	1	0	1	1	0	0	0	1	1	1	1	1	1

Les taux de reconnaissance générés sont satisfaisant par rapport à la forte cursivité et à la variabilité de l'écriture manuscrite arabe d'un scripteur à un autre tel que : la boucle peut être remplisse et ne va pas être détectée, mots avec caractères ligaturés qui seront pris pour un seul, points diacritiques double ou triple fusionnés pris pour un seule …etc.

6 Conclusion

Dans cet article nous avons proposé un algorithme hybride bioinspiré, basé sur les réseaux de neurones comme outil de classification pour la reconnaissance de l'écriture arabe manuscrite et les AG comme optimiseur appliqué sur l'ensemble des attributs caractérisant les mots à reconnaitre. Nous avons appliqué notre algorithme pour la reconnaissance des mots à savoir l'ensemble des villes tunisiennes. Cette application peut présenter beaucoup d'intérêts dans le tri automatique du courrier et la reconnaissance des montants des chèques. Pour généraliser notre contribution pour la reconnaissance d'un texte arabe manuscrit quelconque nous envisageons d'introduire l'aspect sémantique à travers un dictionnaire ou une ontologie.

Références

1. A.Broumandnia and al , "Handwritten Farsi/Arabic Word Recognition",IEEE,(2007).
2. R. Haraty and C. Ghaddar, "Arabic Text Recognition," Int'l Arab J.Information Technology, vol. 1. No 2,.pp 156,163. (2004)
3. M.S. Khorsheed, "Recognising Handwritten Arabic Manuscripts Using a Single Hidden Markov Model,"elseiver Pattern Recognition Letters, vol. 24, (2003).
4. M. Kherallah and al , "On-line Arabic handwriting recognition system based on visuencoding and genetic algorithm", Engineering Applications of Artificial Intelligen, Elsevier, (2008).
5. H.Y. Abdelazim and al, "Arabic text recognition using a partial observation approach", Proc. 12th National Computer Conf., Saudi Arabia, Saudi Arabia, , pp. 427437. 21-24 October (1990)
6. A. Nazif, "A system for the recognition of the printed Arabic characters", Master's Thesis, Faculty of Engineering, Cairo University, (1975).
7. A. Ali Aburas et M. E. Gumah,"'Arabic Handwriting Recognition: Challenges and Solutions"', .IEEE , (2008).
8. A. WT Watkins, "AIRS; A Resource Limited Artificial Immune Classifier".Master thesis, Department of Computer Science. Mississippi State University, (2001).
9. J. Holland. « Adoption in natural and artificial systems ». The MIT press, , 211,(1975).
10. D. Golberg " Genetic Algorithm in Search, Optimisation and Machine L earning". Addison Wesley, (1989).