**Chapitre III :**

**Classification et Apprentissage**

**Automatique**

**Chapitre III**

**Classification et Apprentissage Automatique**

**III .1. Introduction :**

Les classifications manuelles d’une grande quantité des documents textuelle s’avèrent très couteux en temps et en personnel pour cette raison beaucoup de recherche ont été consacré pour le problème de la classification automatique et les travaux portant sur ce sujet date depuis 1960.

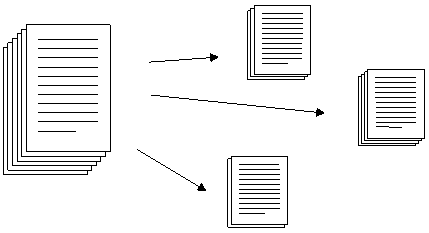
De nos jour beaucoup de logiciels de classification des textes ont fait leur apparition Pour certaines tâches, les classificateurs automatiques performent presque aussi bien que les humains, mais pour d'autres, l'écart est encore grand, ces systèmes sont directement lié à l’apprentissage artificiel (apprentissage automatique ) ce dernier vise à réaliser des taches normalement abordable uniquement par les humains On distingue en général deux types d'apprentissage : l'apprentissage «par cœur» qui consiste à mémoriser telles quelles des informations, et l'apprentissage par généralisation où l'on apprend à partir d'exemples un modèle qui nous permettra de reconnaître de nouveaux exemples.

Dans ce chapitre on va aborder la notion de classification en détaillons comment résoudre les différents problèmes de classification on fait appel aux diffèrent type d’apprentissage automatique.

**III .2-Définition classification des textes :**

La classification de textes est une tâche générique qui consiste à assigner une ou plusieurs catégories, parmi une liste prédéfinie ou non à un document en trouvant une liaison fonctionnelle entre un ensemble de textes et un ensemble de catégories (étiquettes, classes) selon des critères. Elles s'appliquent à un grand nombre d'activités humaines et conviennent en particulier aux problèmes de la prise de décision automatisée.

Par exemple : on est face à un ensemble de texte l'objectif est de rendre une application informatique capable de déterminer de façon autonome, dans quelle catégorie classé chacun des textes, en se basant sur des données statistiques. Habituellement, les catégories font référence aux sujets des textes, mais pour des applications particulières, elles peuvent prendre d'autres formes. **[C03]**



**Figure III.1 classification des textes**

La classification peut être trouvé dans plusieurs domaine d’application comme l’identification de la langue, reconnaissance d’écrivant Catégorisation de document multimédia, Détection de spamme, reconnaissance de la parole, aid aux diagnostique médicale.

**III .3-Classification bi class et multi class :**

**III .3-1-La classification bi-classe** :

La classification bi classe c’est un problème de classification des exemples dans deux classes pour laquelle le système de classification répond à la question : Le texte appartient-il à la catégorie « C » ou à sa catégorie complémentaire. **[C02]**

Certaines tâches typiques de la classification binaires sont :la détermination si un document est-il plagier ou non ,tests médicaux pour déterminer si un patient a certaine maladie ou non , contrôle de la qualité dans les usines décider si un nouveau produit est assez bon pour être vendu, ou si elle doit être éliminée .

**III .3-2-La classification multi-classes :**

La classification multi class est le problème de la classification des exemples dans plus de deux classes ou on peut associer un exemple donné à une ou plusieurs classes voire a aucune classe c’est le cas le plus générale de la classification, qui peut être utilisé dans plusieurs domaine : comme la reconnaissance de caractères manuscrits qui fait partie du discours de marquage, la reconnaissance vocale.

Les méthodes de classification binaires peuvent être généralisés dans de nombreuses façons de gérer plusieurs classes on peut distinguer deux approche pour effectuer cette généralisation. **[C01]**

* **approche** (UCT**) un-contre-tout :** permet l'utilisation directe des classifieurs binaires standard pour encoder et former les étiquettes de sortie et qui suppose que pour chaque classe, il existe une seule (simple) séparation entre cette classe et toutes les autres classes c’est-à-dire que chaque classe est opposer à tous les autres classes. Le Mieme classifieur sépare les données de la classe M de tous les restes de données cette dernière prend signe +1 et les autre classe -1.
* **un contre un (UCU) :** est une alternative plus expressive qui suppose l'existence d'un séparateur entre chaque deux classes cette méthode est dite un contre un c’est-à-dire que chaque classe est discriminée d’une autre ou on doit calculer M(M-1)/2 classificateur . **[C02] [C01]**

## III .4-Pourquoi utilisé la classification automatique des documents?

Les conséquences majeures qui posent un traitement manuel (par un être humain) de classification des documents textuels sont :

* La classification des documents textuel est difficile à être réaliser manuellement par un être humain car elle prend beaucoup de temps pour :
* lire et relire chaque document ligne par ligne et de le mémoriser.
* une grande partie de temps est utilisé dans la réflexion pour la prise de décision finale (associer un texte a une catégorie).
* Le nombre de classe qui jeu un rôle très important dans le cout de la classification.
* La vitesse de lecture change d’une personne a un autre c’est relative donc la prise de décision aussi ne vas pas être la même pour chaque personne.
* Vérifier les autres textes qui sont dans la même catégorie que le texte a classé pour valider la décision.
* La différence entre les classe par exemple classé des document en deux catégorie «médecine» et «droit » est plus facile que classé les document a un des catégories suivant «Traitement automatique des langues » et «Intelligence artificielle»
* La classification doit être faite d’une façon à ce que les groupes obtenus soient constitués d'observations semblables et que les groupes soient le plus différents possible entre eux.
* La généralisation d’une classification manuelle à d’autres domaines est quasi impossible. [C05]

La classification des textes c’est un problème qui peut être lié directement à l’apprentissage automatique (artificiel)

**III .5-L’apprentissage automatique (machine learning) :**

Apprentissage automatique se trouve au carrefour de nombreux autres domaines : intelligence artificielle, statistiques, sciences cognitives, théorie des probabilités, qui fait référence à des algorithmes informatique qui permettent à une machine d’évoluer grâce à un processus d’apprentissage ,on se base sur une sorte d’observation ou des données tel que des exemples ou des expériences pour remplir des tâches qu’il est difficile de les remplir par des moyens algorithmiques plus Classiques comme par exemple faire des prédiction précise ou d’extraire et d’exploiter automatiquement l’information présenter dans un jeu de données . **[C04] [B01]**

En générale : apprentissage automatique c’est d’apprendre à faire mieux à l’avenir sur la base de ce qui a été connu dans le passé ou on peut le considérer comme la programmation par exemple.

«L'apprentissage dénote des changements dans un système qui ... lui permet de faire la même tâche plus efficacement la prochaine fois» ***Herbert Simon****.* L’apprentissage automatique est entourée de plusieurs disciplines comme le montre la figure suivante.

**Figure III.2 discipline autour de l’apprentissage automatique**

**III .5-1-De l’apprentissage natural vers l’artificiel :**

Apprentissage Natural se diffère de l’artificiel le fait qu’il soit réalisé non pas par une machine mais par un humain Par exemple :

Nous demandons à un humain de classer des revue par thème comment il va faire sans connaissance particulier ou préalable ?

Si il est seul il lit les revue et essaye d’observer les points de similitude ou de divergence entre elle donc il peut trouver une façon de regrouper les revue par thème.

Si il est accompagné par un expert peut l’aidé mais le problème qu’il n y’a pas beaucoup de temps à lui consacrer donc il se contente de quelque exemple classé par cet expert pour classé le reste des exemples qui n’a pas eu le temps a les classé. tous sa illustre le processus d’apprentissage artificiel. La machine peut être confrontée à des situation où elle doit apprendre seule (apprentissage non superviser) ou au contraire avec l’aide d’un expert (apprentissage superviser). **[C04]**

**III .5-2-Algorithme d’apprentissage :**

Algorithme d’apprentissage est un algorithme prenant en entrer un ensemble de donné « D » (ensemble d’entrainement ou d’apprentissage) et retourne une fonction **F** (modèle) le modèle obtenue a été entrainer sur l’ensemble « D ».

Ensemble « D » contient sous forme de vecteur l’information nécessaire pour résoudre un problème (par exemple classification).Algorithme d’apprentissage permet l’adaptation automatique à la réalisation d’une tache ou chaque algorithme est connue pour fonctionner sur certain type de tache et pas sûr d’autre.

Le choix d'un algorithme d'apprentissage dépend généralement du type de texte à classifié (formulaire, article, questionnaire, etc.), de l'objectif et du domaine de l'application. Par exemple les applications conçu pour le domaine médicale nécessite des algorithmes d'apprentissage les plus précise que possible qui différent des algorithmes d'apprentissage utilisés pour concevoir un moteur de recherche dans un site web d'une librairie et dépond aussi de la tâche à résoudre (classification, estimation de valeurs, ect) et pour chaque tâche il existe toute une gamme d’algorithmes.

**III .5-3- modèles d’apprentissage**

Le modèle va permettre à l’apprenant d’apprendre efficacement selon un biais pour la réalisation de la tâche désirée. Le choix du modèle à employer est donc primordial pour réussir un apprentissage optimal. Ce biais est d’une nature différente selon le modèle retenu :

* **fonctionnel** : il est alors basé sur une fonction de décision sur les attributs d’entrées permettant d’établir une séparatrice. La plus célèbre et la plus simple est sans doute la fonction linéaire (exemple :SVM support vecteur machine )
* . **probabiliste** : il est alors basé sur une distribution de probabilité sur les attributs des entrées. Les réseaux bayésiens sont sans doute l’exemple typique de ce genre de modèle.
* **connexionniste** : il est alors basé sur un réseau de neurones. Ce type de modèle s’inspire à la base du fonctionnement du cerveau humain Nous citerons les perceptrons et les réseaux de neurone multicouche(MLP) comme exemples typiques connexionnistes.
* **temporelle** : il est alors basé sur un couplage temporelle entre les entrées. Ce type de modèle décrit différents états temporels dans lesquels on est susceptible de se trouver L’exemple typique est le modèle de MARKOV Caché.

**III .5-4-Pourquoi l’apprentissage artificiel**

* Certain tâche ne sont bien définie que via un ensemble d’exemple pour bien spécifier les relations entre les entrées sortie
* La quantité des données peut contenir des relations importantes que les méthodes d’apprentissage artificiel permettre de le découvrir.
* L’apprentissage artificiel peut permettre de mieux exploiter les connaissances ou le cerveau humain ne puisse les explicité.
* L’apprentissage artificiel permet aux machines de s’adapter aux changements de l’environnement sans refaire systématiquement la conception après chaque changement de l'environnement

**III .5-5-les types d’apprentissages :**

**III .5-5-1-apprentissage supervisé :**

Un superviseur (expert ou oracle) est disponible pour étiqueté correctement les données d’apprentissage ou d’entrainement dans ce cas, D correspond à un ensemble de « n » paires d’entrées « xt » et d’étiquette associées « yt » et la fonction « F » vas être définie d’une façon explicite par l’algorithme d’apprentissage ce type d’apprentissage est constitué de deux phase

* Phase d’apprentissage (d’entrainement) : dans cette phase l’algorithme d’apprentissage reçoit en entrer des exemple d’apprentissage étiqueté « D » et produit un modèle de prédiction le plus performant possible, c’est-à-dire le modèle qui produit le moins d’erreurs en prédiction comme présente la **figure III.3**.

**Prédiction** : cette tache vise à prédire une ou plusieurs caractéristiques inconnues à partir d’un ensemble de caractéristique connues .Par exemple :prédire la qualité d’un client en fonction de son revenue et de son nombre d’enfant  **.[C05]**

Les documents

D’entrainement

indexation

Les class prédéfini

Model de prédiction

**Figure III.3 phase d’apprentisage**

* **la phase de test** : c’est la phase de prédiction des nouveaux exemple en fonction du modèle obtenue lors de la phase d’apprentissage qui vas associer à chaque entrer « X » quelconque une sortie (étiquette) « Y » qui auras normalement été donnée par le superviseur (tester la précision du modèle).

**Phase de classification (test)**

Indexation

Nouveaux document test

Model de prédiction

Texte catégoriser

**Figure III.4 phase de test**

Ces deux phases peuvent être exécutées de manières consécutive apprentissage supervisé hors ligne ou itérative apprentissage supervisé en ligne.

**III .5-5-1-1-problème résolu par apprentissage superviser :**

l’application de l’apprentissage superviser concerne la médicine : étant données les résultats d’analyse d’un patient et la connaissance de l’état d’autre patients pour lesquels les même analyse ont été menées il est possible d’évaluer le risque de maladie de ce nouveau patient en fonction de la similarité de ses analyses avec celles des autre patient .

On distingue en général deux grand types de problèmes auxquels l'apprentissage supervisé est appliqué : la classification supervisée (catégorisation), la régression .

**Catégorisation de texte** (classification supervisé) : Elle permet de prédire si un texte est membre d’un groupe ou d’une classe prédéfinie .C’est la classification de chaque texte x ∈ X parmi un ensemble de classe préexistantes (connu à l’ avance)

Le principe générale c’est de chercher un modèle de prédiction estimer par un apprentissage supervisé .Ce typer de classification peut être utiliser dans plusieurs domaine par exemple

Filtrage anti-spam: identifier les e-mails comme spam ou non-spam

Diagnostic médical: diagnostiquer un patient comme un malade ou non malade d'une maladie

Prévision météo: prédire, par exemple : si oui ou non il va pleuvoir demain.

**III .5-5-2-Apprentissage non supervisé** :

Aucun superviseur n’est disponible et les données d’apprentissage **D** contient que des entrées « X » non étiquette et le modèle de prédiction produit par l’algorithme d’apprentissage seras implicite. L’algorithme doit découvrir par lui-même les similitudes et divergences entre les données et vise à caractériser la distribution des données, et les relations entre les variables, sans discriminer entre les variables observées et les variables à prédire.

pour ce type d’apprentissage c’est plutôt l’utilisateur qui doit spécifier le problème à résoudre .la **figure III.5** présente un apprenant qui reçoit trois entrèe X1,X2,X3 et en sortie X1 et X2 sont étiquette par Y1 et X3 par Y2.



**Figure III.5 fonctionnement d’apprentissage non supervisée**

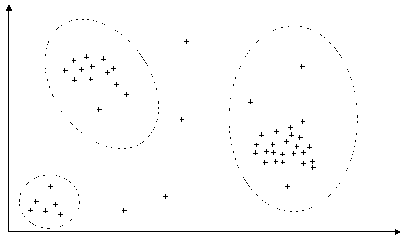
**III .5-5-2-1-Problème résolu par apprentissage non supervisée :**

L’apprentissage non superviser peut être adapter a des problème comme : Extraction de caractéristiques, Réduction de dimensionnalité, clustering.

Classification non supervisée (clustering) : le problème de clustering c’est un problème de Partitionnement logique de l’ensemble des données D a des sous ensemble (cluster) ou chaque cluster possède un certain degré de cohérence interne comme le montre la **Figure III .6**.

La classification non superviser est la recherche des classes principale de la distribution par le processus de l’apprentissage non superviser puisque les classe ne sont pas connue à l’avance(donnée cache ).

De même on peut aussi soit considérer que chaque exemple ne peut appartenir qu’a un seul groupe (hard clustering) soit considérer que chaque exemple a une probabilité donnée d’appartenir à chacun des groups (soft clustering ) . **[C03] [C01]**



**Figure III .6 clustering**

Le problème est ici plus difficile puisque les observations disponibles ne sont pas initialement identifiées comme appartenant à telle ou telle population : il faudra déduire cette information lors de notre classification.

**Exemple d’application :**

Marketing : segmentation du marché en découvrant des groups de clients distincts à partir de bases de données d’achats.

Planification de villes: identification de groupes d’habitations suivant le type d’habitation, valeur, localisation géographique ……ect

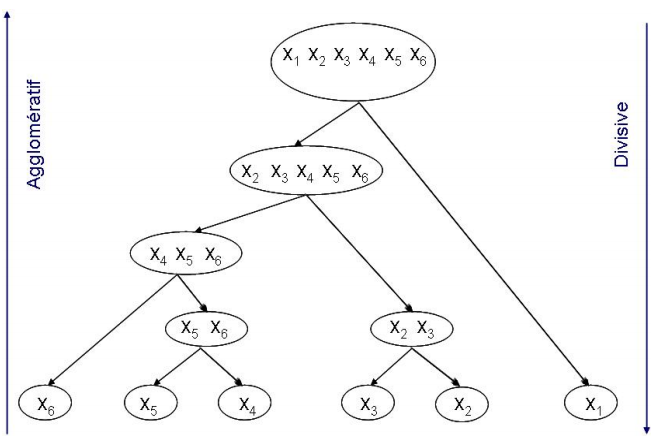
Assurance: identification de groupes d’assurés distincts associés à un nombre important de déclarations.

**III .5-6-Les algorithme d’apprentissage non supervisée :**

**III .5-6-1-La classification hiérarchique :**

Permet de construire un arbre de clusters, appelé dendrogramme, qui montre comment les clusters sont organisés .ou le nombre de cluster « K » n’est pas exigé comme donnée En coupant le dendrogramme au niveau désiré, une classification des données dans des groupes disjoints est obtenue avec l’utilisation d’une matrice de distance comme critère de clustering . **[C06]**

Les méthodes de classification hiérarchique sont divisées en deux type d’approches : ascendante et descendante qui sont présentées dans la figure suivante :



**Figure III.7 classification ascendante et descendante**

**III .5-6-2-Les algorithme de partitionnement des données(**non hiérarchique)**:**

Ces algorithme partage les données en plusieurs sous ensemble et cherche à décomposer directement la base de données dans un ensemble de cluster disjoints Plus spécifiquement, ils essaient de déterminer un nombre de partitions qui optimisent certains critères (fonction objective). [C04] La fonction objective peut mettre en valeur la structure locale ou globale des données et son optimisation est une procédure itérative Parmi les algorithmes non hiérarchique le plus utilisè :

**III .5-6-2-1-kmeans (k-moyenne) :**

K-means à l’ intention de partage n objets en k classes où chaque objet appartient au cluster avec le plus proche moyenne. Cette méthode produit exactement k différents groupes de distinction le plus grand possible. Le meilleur nombre de clusters k conduisant à la plus grande séparation (distance) n'est pas connu comme a priori et doivent être calculées à partir des données. L'objectif de K-means est de minimiser les échanges totaux intra-cluster variance, ou la fonction de l'erreur quadratique.

|  |
| --- |
|  |
|  |
|  |

**Algorithme**

Entrée: un échantillon de données x1, …, xm

1- Sélectionnez k points au hasard, comme les centres de classes.c1, …, ck

2- Assigner chacun des données m dans le groupe i dont le centre ci est le plus proche. en utilisant la fonction de distance Euclidienne

3-Si aucun élément ne change de groupe alors arrêt et sortir les groupes

4-Calculer les nouveaux centres : pour tout i, ci est la moyenne des éléments du groupe i.

5- Aller en 2 et répéter les étapes jusqu'à ce que les mêmes points sont attribués à chaque groupe

**Les avantages :**

* Relativement extensible dans le traitement d’ensemble de taille importante
* Les résultats obtenus sont relativement efficace
* Produit généralement un optimum local et un optimum global peut être obtenu en utilisant d’autre technique : les algorithmes génétique ….ect
* Les inconvénient : applicable seulement dans le cas où la moyenne des objets est définie ,besoin de spécifier K le nombre de cluster a priori incapable de traiter les données bruité ,les point isoler sont mal gérés.

**III .5-6-3-Méthode basées sur la densité** :

l’utilisation de mesures de similarité (distance) est moins efficace que l’utilisation de densité de voisinage idée centrale est de grouper des objets voisins de l'ensemble des données en classes suivant une fonction de densité parce que la minimisation de la distance inter-clusters n’est pas toujours un bon critère pour reconnaitre des « forme ».

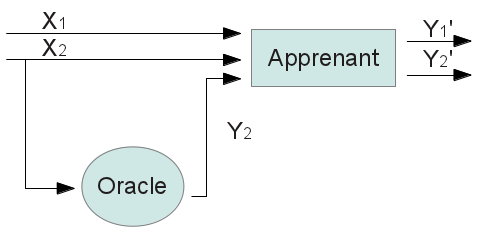
**III .5-7-L’apprentissage semi superviser (hybride) :**

C’est un mélange entre le principe de base de l’apprentissage superviser et non superviser utiliser lorsque l’étiquetage des exemple est couteux ce type d’apprentissage qui a connu beaucoup d’importance au cours des dernières années a permet d’améliorer les résultat d’apprentissage **.**

Un superviseur est partiellement disponible et étiquette quelques données. L’algorithme va alors apprendre la tâche de classification en se basant sur les étiquettes posées par le superviseur et en découvrant par lui-même les informations manquantes (les exemple non étiqueté).

Les algorithmes semi-supervisés fonctionnent sur les deux même phases mais acceptent en plus des données non étiquetées pendant la phase d’entraînement comme la figure suivante ou l’ensemble d’apprentissage (x1 non étiqueté et x2 étiqueté ) . **[C05]**

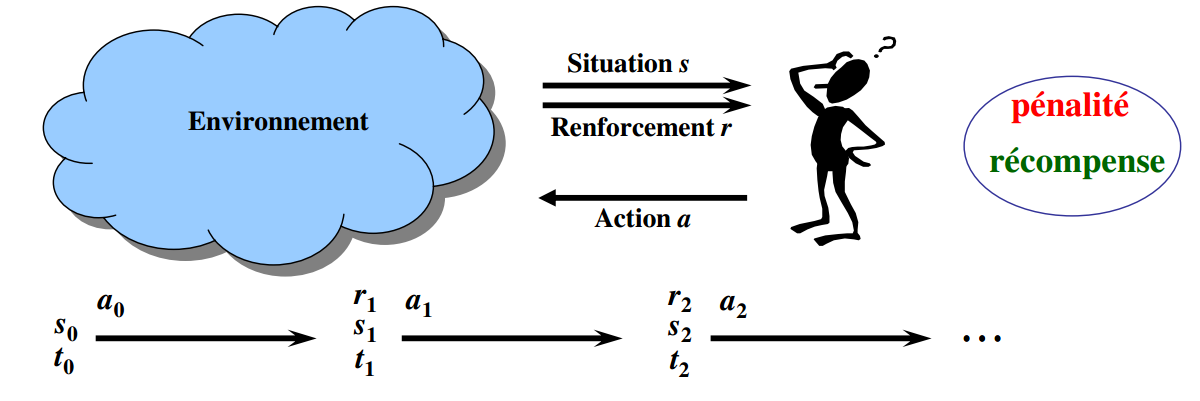
Ex : En médecine, il peut constituer une aide au diagnostic.



**Figure III.8 fonctionnement d’apprentissage semi supervisé**

**III 5-8-Apprentissage par renforcement :**

L’objectif principale c’est d’entraîner un agent à se comporter de façon intelligente dans un environnement donné. Un agent interagit avec l’environnement en choisissant, à chaque temps donné, d’exécuter une action parmi un ensemble d’actions permises. Le comportement intelligent que doit apprendre cet agent est donné implicitement via un signal de renforcement qui, après chaque décision de l’agent, indique s’il a bien ou mal agi et l’agent a comme entrer un ensemble d’indicateur ou de caractéristique décrivant l’environnement comme le montre la figure suivante :



**Figure III.9 fonctionnement d’apprentissage par renforcement**

Ex : L'algorithme de [Q-learning](http://fr.wikipedia.org/w/index.php?title=Q-learning&action=edit&redlink=1)(la notion de temps n’ existe pas) ou Td-Learning (dynamique :la notion de temps existe). Ce type d'apprentissage est appliqué dans de nombreux domaines comme les jeux (dames), la robotique…. ect

**Conclusion**

Ce chapitre englobe le problème de classification de texte et comment résoudre ce type de problème à l’aïd d’un apprentissage automatique on a vue aussi les défirent type d’apprentissage leur limites et les domaines d’application de chacune d’eux en détaillons quelque algorithme d’apprentissage les plus populaires.