Chapitre II

Data mining

II.1 Introduction

Les techniques de data mining sont utilisé de façon augmentaté dans le domaine économique. Tels que la prédiction de certains indicateurs économiques, la découverte des informations cachées, des problèmes ou de trouver des problèmes dans le secteur industriel, ainsi que dans les relations avec les clients à travers l'étude de leurs données et leurs comportements afin d'améliorer le rapport coût-efficacité de la relation avec les clients ou d'attirer de nouveaux clients.

Dans ce chapitre nous voulons reconnaitre les différentes techniques de data mining afin d'avoir un aperçu complet sur eux, pour identifier les technique appropriées pour l'utilisé dans la résolution des problèmes trouvé dans la premier chapitre.

II.2 Définition du data mining

Data mining signifie l'extraction de connaissance a travers l'analyse d'une grande quantité de données pour utiliser ces connaissances dans le processus de décision [TSI09]. On peut trouver les données stocker et organisé dans les datamarts et les entrepôts de données, ou dans autre sources non structurées.

Le processus de data mining implique plusieurs étapes avant de trouver un modèle de décision qui peut être un ensemble de règles, des équation ou des fonction de transfert complexes. Selon leur objectif le data mining se compose en deux catégories supervisé et non supervisé, qui nous en parler dans les technique de data mining.

II.3 Processus de data mining [DJE13]

Il est très important de comprendre que le data mining n'est pas seulement le problème de découverte de modèles dans un ensemble de donnée. Ce n'est qu'une seule étape dans tout un processus suivi par les scientifiques, les ingénieurs ou toute autre personne qui cherche à extraire les connaissances à partir des données. En 1996 un groupe d'analystes définit le data mining comme étant un processus composé de cinq étapes sous le standard CRISP-DM (Cross-Industry Standard Process for Data Mining) comme schématisé ci-dessous (figure 2.1) :

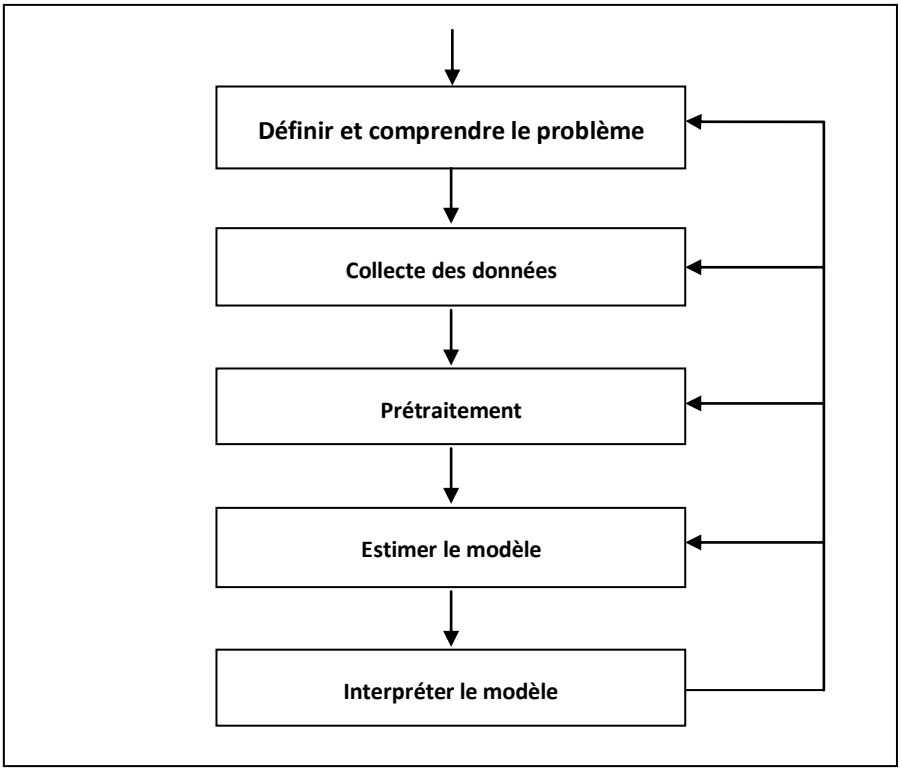

Figure 2.1 : Processus de data mining

Définition et compréhension du problème : Dans la plus part des cas, il est indispensable de comprendre la signification des données et le domaine à explorer. Sans cette compréhension, aucun algorithme ne va donner un résultat fiable. En effet, Avec la compréhension du problème, on peut préparer les données nécessaires à l'exploration et interpréter correctement les résultats obtenus.

Collecte des données : D'après la définition du problème et des objectifs du data mining, on peut avoir une idée sur les données qui doivent être utilisées. Ces données n'ont pas toujours le même format et la même structure. On peut avoir des textes, des bases de données, des pages web, ...etc. Parfois, on est amené à prendre une copie d'un système d'information en cours d'exécution, puis ramasser les données de sources éventuellement hétérogènes (fichiers, bases de données relationnelles, temporelles, ...).

Prétraitement : Les données peuvent contenir plusieurs types d'anomalies : des données peuvent être omises à cause des erreurs de frappe ou à causes des erreurs dues au système lui-même, dans ce cas il faut remplacer ces données ou éliminer complètement leurs enregistrements.

Des données peuvent être incohérentes c-à-d qui sortent des intervalles permis, on doit les écarter où les normaliser. Parfois on est obligé à faire des transformations sur les données pour unifier leur poids. Le prétraitement comporte aussi la réduction des données qui permet de réduire le nombre d'attributs pour accélérer les calculs et représenter les données sous un format optimal pour l'exploration. Dans la majorité

des cas, le pré-traitement doit préparer des informations globales sur les données pour les étapes qui suivent tel que la tendance centrale des données (moyenne, médiane, mode), le maximum et le minimum, le rang, les quartiles, la variance, ... etc.

Plusieurs techniques de visualisation des données telles que les courbes, les diagrammes, les graphes,... etc, peuvent aider à la sélection et le nettoyage des données. Une fois les données collectées, nettoyées et prétraitées on les appelle entrepôt de données (data warehouse).

Estimation du modèle : Dans cette étape, on doit choisir la bonne technique pour extraire les connaissances (exploration) des données. Des techniques telles que les réseaux de neurones, les arbres de décision, les réseaux bayésiens, le clustering, ... sont utilisées. Généralement, l'implémentation se base sur plusieurs de ces techniques, puis on choisit le bon résultat. Dans le titre suivant on va détailler les différentes techniques utilisées dans l'exploration des données et l'estimation du modèle.

Interprétation du modèle et établissement des conclusions : généralement, l'objectif du data mining est d'aider à la prise de décision en fournissant des modèles compréhensibles aux utilisateurs. En effet, les utilisateurs ne demandent pas des pages et des pages de chiffres, mais des interprétations des modèles obtenus. Les expériences montrent que les modèles simples sont plus compréhensibles mais moins précis, alors que ceux complexes sont plus précis mais difficiles à interpréter.

II.4 Techniques du data mining

II.4.1 Techniques supervisées [TSI09]

Dans la modélisation supervisée, ou prédictive, l'objectif est de prédire un événement ou d'estimer les valeurs d'un attribut numérique continue. Dans ces modèles, il existe des champs où les attributs d'entrée et une zone de sortie ou de la cible. Les champs d'entrée sont également appelés prédicteurs, car ils sont utilisés par le modèle pour identifier une fonction de prédiction de champ de sortie. Nous pouvons penser à des prédicteurs que la partie X de la fonction et le domaine cible que la partie Y, le résultat.

Le modèle utilise les champs de saisie qui sont analysées en ce qui concerne leur effet sur le champ cible. La reconnaissance de formes est "surveillé" par le domaine cible. Des relations sont établies entre les champs d'entrée et de sortie. Une cartographie " fonction d'entrée-sortie " est généré par le modèle, qui associe des prédicteurs et à la sortie permet la prédiction des valeurs de sortie, étant donné les valeurs des champs d'entrée.

Les modèles prédictifs sont subdivisés en modèles de classification et d'estimation :

• **Les modèle classification:** Dans ces modèles les groupes ou classes cibles sont connus dès le départ. Le but est de classer les cas dans ces groupes prédéfinis ; en d'autres termes, à prévoir un événement. Le modèle généré peut être utilisé comme un moteur de marquage pour l'affectation de nouveaux cas pour les classes prédéfinies. Il estime aussi un score de propension pour chaque cas. Le score de propension dénote la probabilité d'occurrence du groupe cible ou d'un événement.

• **Les modèle d'estimation :** Ces modèles sont similaires à des modèles de classification, mais avec une différence majeure. Ils sont utilisés pour prédire la valeur d'un champ continu en fonction des valeurs observées des attributs d'entrée.

II.4.1.1 Arbre de décision

Les arbres de décision fonctionnent en séparant de façon récursive la population initiale. Pour chaque groupe, ils sélectionnent automatiquement l'indicateur le plus significatif, le prédicteur qui donne la meilleure séparation par rapport au champ cible. À travers des cloisons successives, leur objectif est de produire sous-segments pures, avec un comportement homogène en termes de production. Ils sont peut-être la technique la plus populaire de classification. Une partie de leur popularité, c'est parce qu'ils produisent des résultats transparents qui sont facilement interprétables, offrant un aperçu de l'événement à l'étude. Les résultats obtenus peuvent avoir deux formats équivalents. Dans un format de règle, les résultats sont représentés dans un langage simple que les règles ordinaires :

SI (VALEURS PREDICTIVES) ALORS (RESULTAT CIBLE ET SCORE DE CONFIANCE).

Dans une forme d'arborescence, les règles sont représentés graphiquement sous forme d'arbre dans laquelle la population initiale (nœud racine) est successivement divisé en des nœuds terminaux ou feuilles de sous-segments ayant un comportement similaire en ce qui concerne le champ cible.

Les algorithmes d'arbres de décision constituent selon la vitesse et l'évolutivité. Algorithmes disponibles sont:

- C5.0

- CHAID

- Classification et arbres de régression

- QUEST.

II.4.1.2 Règles de décision

Ils sont assez semblables à des arbres de décision et de produire une liste de règles qui ont le format des états humains compréhensible:

SI (VALEURS PREDICTIVES) ALORS (RESULTAT CIBLE ET SCORE DE CONFIANCE).

Leur principale différence par arbres de décision, c'est qu'ils peuvent produire plusieurs règles pour chaque enregistrement. Les arbres de décision génèrent des règles exhaustives et mutuellement exclusifs qui couvrent tous les records. Pour chaque enregistrement une seule règle s'applique. Au contraire, les règles de décision peuvent générer un ensemble de règles de chevauchement. Plus d'une règle, avec des prédictions différentes, peut être vrai pour chaque enregistrement. Dans ce cas, les règles sont évaluées, à travers une procédure intégrée, afin de déterminer l'une pour l'évaluation. Habituellement, une procédure de vote est appliquée, qui combine les règles et les moyennes de leurs confidences individuelles pour chaque catégorie de sortie. Enfin, la catégorie ayant la confiance la moyenne la plus élevée est sélectionnée comme la prédiction.

Les algorithmes de règles de décision comprennent :

- C5.0

- Liste de décision.

II.4.1.3 Régression

La régression est la méthode utilisée pour l'estimation des valeurs continues. Son objectif est de trouver le meilleur modèle qui décrit la relation entre une variable continue de sortie et une ou plusieurs variables d'entrée. Il s'agit de trouver une fonction f qui se rapproche le plus possible d'un scénario donné d'entrées et de sorties [DJE13].

II.4.1.4 Réseaux de neurone

Les réseaux de neurones sont des puissants algorithmes d'apprentissage automatique qui utilisent des fonctions de cartographie complexe, non linéaire pour l'estimation et classification. Ils sont constitués de neurones organisés en couches. La couche d'entrée contient les prédicteurs ou neurones d'entrée. La couche de sortie comprend dans le champ cible. Ces modèles permettent d'estimer des poids qui relient les prédicteurs (couche d'entrée à la sortie). Modèles avec des topologies plus complexes peuvent également inclure, couches cachées intermédiaires, et les neurones. La procédure de formation est un processus itératif. Enregistrements en entrée, avec des résultats connus, sont présentés sur le réseau et la prédiction du

modèle est évaluée par rapport aux résultats observés. Erreurs observées sont utilisés pour ajuster et d'optimiser les estimations du poids initial. Ils sont considérés comme des solutions opaques ou "boîte noir" car ils ne fournissent pas une explication de leurs prédictions. Ils fournissent seulement une analyse de sensibilité, qui résume l'importance prédictive des champs d'entrée. Ils nécessitent une connaissance statistique minimum mais, selon le problème, peut nécessiter un temps de traitement à long pour la formation.

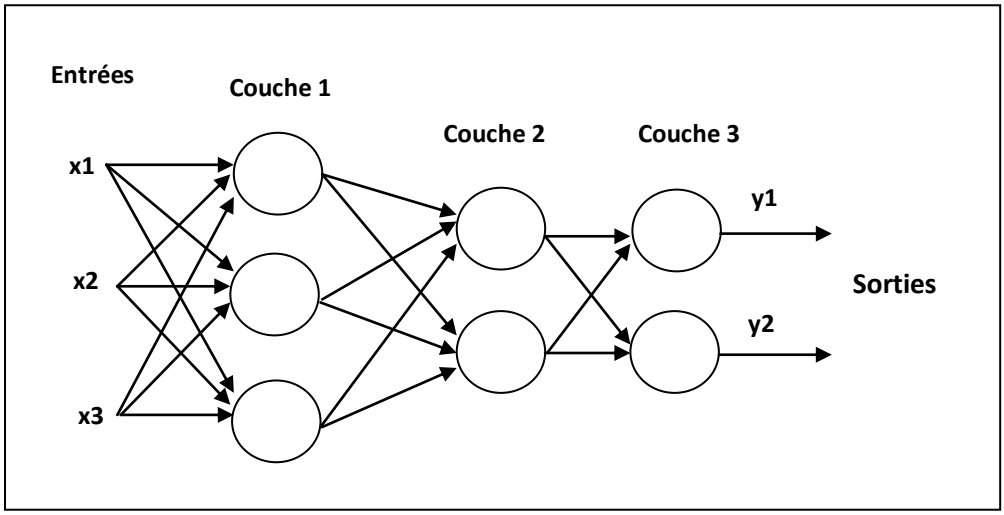

Figure 2.2 : Réseau de neurones artificiel

II.4.1.5 Machines à vecteurs supports (SVM)

SVM est un algorithme de classification qui peut modéliser les profils de données non linéaires hautement complexes, et d'éviter les sur-apprentissages, c'est-à-dire la situation dans laquelle un modèle mémorise les modèles ne concernent que des cas spécifiques analysés. SVM fonctionne en données cartographiques à un espace de grande dimension caractéristique dans lequel les enregistrements deviennent plus facilement séparables (ie, séparés par des fonctions linéaires) à l'égard des catégories de cibles.

Les données d'entraînement d'entrée sont transformés de manière appropriée par les fonctions du noyau non linéaires et cette transformation est suivie d'une recherche de fonctions plus simples, c'est-à des fonctions linéaires, qui enregistre de façon optimale distincts. Les analystes expérimentent généralement avec différentes fonctions de transformation et de comparer les résultats. Globalement SVM est un algorithme efficace et exigeant, en termes de ressources de mémoire et de temps de traitement. En outre, il manque de transparence puisque les prévisions ne sont pas expliquées et seulement l'importance des prédicteurs est résumée.

II.4.1.6 Réseaux bayésiens

Les modèles bayésiens sont des modèles de probabilité qui peuvent être utilisées dans des problèmes de classification pour estimer la probabilité d'occurrences. Ils sont des modèles graphiques qui fournissent une représentation visuelle des relations d'attributs, en assurant la transparence, et une explication de la justification du modèle.

II.4.2 Techniques non supervisées

Dans les modèles non supervisés ou non orientés, il n'y a pas de champ de sortie, il n'y a que des entrées. La reconnaissance de formes est non orienté; elle n'est pas guidée par un attribut cible spécifique. Le but de ces modèles est de découvrir des motifs de données dans l'ensemble des champs d'entrée.

Les modèles non supervisés comprennent :

• **Les modèles de dispersion :** Dans ces modèles les groupes ne sont pas connus à l'avance. Au contraire, nous voulons que les algorithmes pour analyser les schémas de données d'entrée et d'identifier les regroupements naturels de données ou de cas. Lorsque de nouveaux cas sont marqués par le modèle de cluster généré ils sont affectés à l'un des groupes révélés.

• **Les modèles d'association de séquences :** Ces modèles font également partie de la classe de la modélisation non supervisé. Ils ne comportent pas de prédiction directe d'un seul champ. En fait, tous les champs concernés ont un double rôle, car ils agissent comme des entrées et des sorties en même temps. Des modèles d'association de détecter des associations entre des événements discrets, des produits ou des attributs. Les modèles de séquence détectent des associations au fil du temps.

II.4.2.1 Clustring hiérarchique

Il considère comme la "mère" de tous les modèles de clustering. Il est appelé hiérarchique ou d'agglomération, car il commence avec une solution où chaque enregistrement comprend un groupe et peu à peu les groupes se former jusqu'au point où tous tomber dans un super-cluster (figure 2.3). À chaque étape, il calcule les distances entre toutes les paires d'enregistrements et les groupes les plus similaires. Une table (horaire d'agglomération) ou un graphique (dendrogramme) résume les étapes de regroupement et les distances respectives.

L'analyste doit consulter ces informations, identifier le point où l'algorithme commence à cas disjoints de groupe, et de décider ensuite sur le nombre de grappes à conserver. Cet algorithme ne peut pas traiter efficacement plus de quelques milliers de cas. Ainsi, il ne peut pas être directement appliqué dans la plupart des tâches de regroupement d'entreprise. Une solution habituelle consiste à une utilisation sur un

échantillon de la population de clustering. Cependant, de nombreux autres algorithmes efficaces qui peuvent facilement gérer des millions d'enregistrements, le regroupement par échantillonnage n'est pas considéré comme une approche idéale.

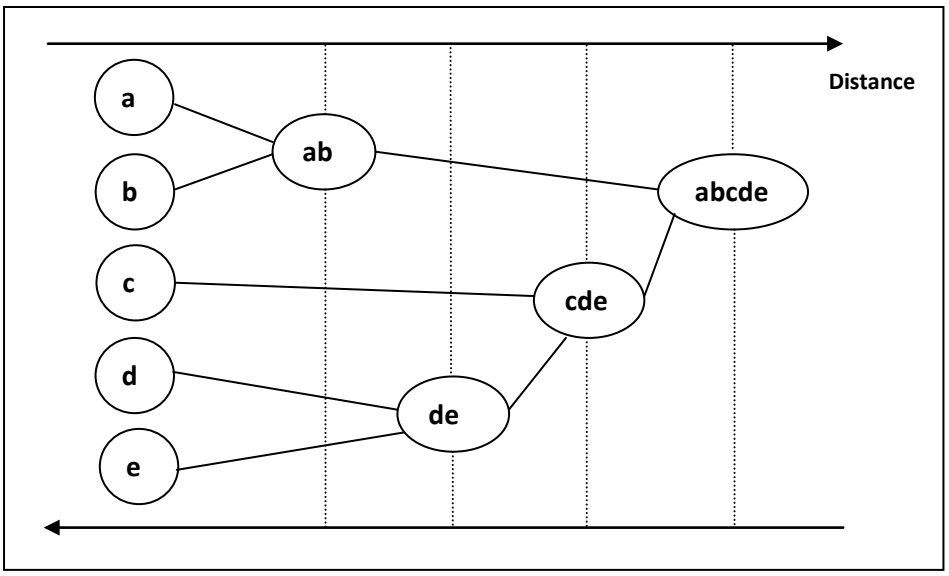

Figure 2.3 : Clustring hiérarchique [KAU01]

II.4.2.2 K-means

C'est un moyen efficace et peut-être l'algorithme de segmentation le plus rapide qui peut gérer deux longues (plusieurs enregistrements) et des ensembles de données larges (de nombreuses dimensions de données et des champs d'entrée). Il s'agit d'une technique de segmentatin basé sur la distance et, à la différence de l'algorithme hiérarchique, il n'a pas besoin de calculer les distances entre toutes les paires d'enregistrements. Le nombre de grappes d'être formés et est prédéterminée spécifiée par l'utilisateur à l'avance. Habituellement, un certain nombre de solutions différentes doit être jugé et évalué avant d'approuver le plus approprié.

II.4.2.3 Carte auto-organisatrice de Kohonen

Réseaux de Kohonen sont basés sur des réseaux de neuronaux et produisent typiquement une grille à deux dimensions ou une carte des grappes, où les cartes d'auto-organisation. Réseaux de Kohonen prennent généralement plus de temps à former que les K-means, mais ils fournissent un point de vue différent sur le regroupement qui est la peine d'essayer.

II.5 Conclusion

Nous avons fait un survol sur les techniques de data mining. Et nous avons constaté qu'ils sont décomposés en deux catégories, la première se compose des techniques supervisées et la deuxième se compose des techniques non supervisé. Dans ces deux catégories, il existe de nombreuses techniques avec des caractéristiques différentes et avec des points fortes et d'autres faibles. Nous allons utiliser ces informations pour choisir les techniques nécessaires pour résoudre les problèmes que nous avons mentionnés dans le premier chapitre.