Table des matières

CHAPITRE I: les séries temporelles et la classification.

LISTE DES FIGURES

CHAPITRE I: les séries temporelles et la classification.

CHAPITRE II: Segmentation des séries temporelles.

CHAPITRE III: Conception et Implémentation.

Liste des tableaux

Introduction Générale

La fouille de données est un domaine pluridisciplinaire permettant à partir d'une énorme masse de données d'extraire de façon automatique ou semi-automatique des informations cachées, pertinentes et inconnues auparavant en vue d'une utilisation industrielle ou opérationnelle. Il désigne l'ensemble des méthodes destinées à l'exploration et à l'analyse de grandes bases de données en vue de détecter dans ces données des profils-type, des comportements récurrents, des règles, des tendances inconnues, des structures particulières restituant de façon concise l'essentiel de l'information utile pour l'aide à la décision.

Parmi les types de données les plus importants que les chercheurs de ce domaine se sont intéressés et ont consacré leur temps et leurs efforts à les comprendre et à les analyser, on trouve les données temporelles, bien connues sous le nom de séries temporelles ou celui de séries chronologiques. Il s'agit d'ensembles de valeurs numériques observées, relatives à un phénomène qui évolue dans le temps. Elles peuvent s'agir de données macroéconomiques telles que le PIB d'un pays, l'inflation, ou les exportations, microéconomiques comme les ventes d'une entreprise ou le revenu d'un individu, financières comme le cours d'une action, politiques comme le nombre de votants ou de voix reçues par un candidat, ou enfin, démographiques comme la taille moyenne des habitants ou leurs âges.

Dans ce travail nous allons nous intéressés à la classification non supervisée des séries temporelles du fait que c'est un type de données omniprésent dans beaucoup de domaines d'application et beaucoup de recherches y se sont consacrées. L'objectif de notre travail est d'étudier les techniques utilisées pour cette classification, connue aussi sous le nom de segmentation ou de regroupement, et de concevoir un outil à ce fait. Pour ce fait, dans le premier chapitre, nous allons introduire le domaine de la fouille de données et définir la tâche de classification et son principe, et nous nous approfondirons dans la définition des séries temporelles, et les méthodes les plus courantes pour leur analyse. Le deuxième chapitre sera consacré à la classification non supervisée des séries temporelles, ses quatre composantes et les méthodes utilisées dans chaque composante. Le troisième chapitre contiendra les détails de conception et d'implémentation de notre outil, avec le déroulement d'un exemple illustratif. Enfin, nous terminerons par une conclusion tout en exposant certaines perspectives.

Chapitre I :
Les séries temporelles et la classification

I.1 Introduction

La notion de classification est essentielle en science en terme générale car c'est une étape de base de traitement et elle permet aux scientifiques de mettre de l'ordre dans les connaissances qu'ils ont sur le monde. Ainsi, depuis longtemps des nombreuses classifications ont été créées par des chercheurs et des scientifiques dans plusieurs secteurs comme le secteur industriel, médical, scientifique, juridique, géographique, économique,….etc. La classification est considérée comme la base de tout processus de recherche, et il existe plusieurs méthodes pour la réaliser. Cette diversité dépend de l'application visée. Ces méthodes peuvent se regroupées en deux grandes catégorie : les méthodes de classification supervisée et les méthodes de classification non supervisée. Ce qui nous concerne au premier lieu dans ce travail est la classification non supervisée, et notre objectif est d'arriver à classer, non pas des objets représentés pas des valeurs, mais des séries de valeurs ou séries temporelles. Une série temporelle est une collection de données obtenues de manière séquentielle au cours du temps.

I.2 La fouille de données ou le data mining

I.2.1 Définition :

Le data mining est le processus de découverte significative de nouvelles corrélations, tendances et caractéristiques dans de grandes quantités de données stockées dans des entrepôts, en utilisant les technologies de reconnaissance de formes, des statistiques, et les techniques mathématiques. Le data mining peut aussi être défini comme « l'analyse de données d'observation fixes afin de trouver des relations insoupçonnées et représenter ces données de façon originale et compréhensible pour le preneur de décision ». Il peut également être définit comme un « domaine interdisciplinaire qui utilise des techniques d'apprentissage automatique, de la reconnaissance des formes, des statistiques, des bases de données et de la visualisation pour l'extraction d'informations à partir de bases de données volumineuses ».

Selon certaines magazines en ligne des nouvelles technologies, le data mining est « l'une des sciences les plus révolutionnaires de la prochaine décennie » [1]. Il sert à trouver des structures originales et des corrélations informelles entre les données. Il permet de mieux comprendre les liens entre des phénomènes en apparence distincts et d'anticiper des tendances encore peu discernables [2]. Donc à partir de ces différent définitions en arrive à ce concept : « le data mining veux dire l'extraction d'informations intéressantes (non triviales, implicites, préalablement inconnues et potentiellement utiles) à partir de grandes bases de données ».

I.2.2 Domaine d'application :

Il existe plusieurs domaines d'application du data mining parmi lesquels :

La gestion et l'analyse de risque :

Les applications de gestion de risque utilisent data mining pour déterminer les primes d'assurances, la gestion de portefeuilles d'investissements, pour différencier entre les entreprises et/ou particuliers qui sont « bons / mauvais » du point de vue de risque de crédit.

La santé :

Prédire si un patient, hospitalisé en raison d'un infarctus, aura une deuxième crise cardiaque. Une prédiction peut se baser sur les données démographiques, régime alimentaire, des mesures cliniques, etc. ou identifier les facteurs de risque de cancer de la prostate à partir des variables cliniques et démographiques, ou encore faire le choix du médicament le plus approprié pour guérir une maladie donné.

Détection de fraude :

Le « Financial Crimes En forcement Network AI Systèmes » (SIAF) utilise des technologies de data mining pour identifier les activités de blanchiment d'argent possibles au sein d'opérations importantes en espèces [3].

Marketing direct :

Population à cibler (âge, sexe, profession, habitation….)

Bioinformatique et génome :

ADN mining, Analyse de génome, mise au point de médicament.

I.2.3 Pourquoi le data mining :

Il existe trois intérêts essentiel pour le data mining à atteindre et sont comme suit :

Expliquer :

Dans cette phase le chercheur doit interpréter un phénomène après un diagnostic, et après la consultation des informations contenues dans un entrepôt de données.

Confirmer :

Le data mining nous aide à confirmer un comportement ou une hypothèse à l'aide des méthodes statistiques ou d'intelligence artificielle.

Explorer :

Le data mining peut explorer les données pour découvrir des liens, ce qui nous permet de suggérer des hypothèses. La décision finale appartiendra toujours au décideur.

I.3 Principe du data mining :

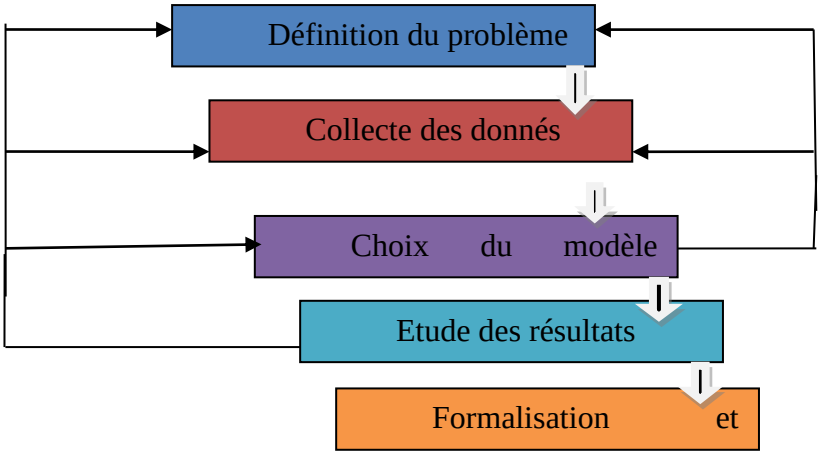

Figure I. 1.Principe du data mining.

Définition du problème :

Quel est le but de l'analyse, que recherche-t-on ? Quels sont les objectifs ? Comment traduire le problème en une question pouvant servir de sujet d'enquête pour cet outil d'analyse bien spécifique ? A ce sujet, se souvenir que l'on travaille à partir des données existantes, la question doit être ciblée selon les données disponibles.

Collecte des données :

Une phase absolument essentielle où on n'analyse que des données utilisables, c'est à dire « propres » et consolidées. On n'hésitera pas à extraire de l'analyse les données de qualité douteuse. Bien souvent, les données méritent d'être retravaillées. S'assurer au final que la quantité de données soit suffisante pour éviter de fausser les résultats. Cette phase de collecte nécessite le plus grand soin.

Construire le modèle d'analyse :

Ne pas hésiter à valider le choix d'analyse sur plusieurs jeux d'essais en variant les échantillons. Une première évaluation peut nous conduire à reprendre le point 1 ou 2.

Etude des résultats :

Il est temps d'exploiter les résultats. Pour affiner l'analyse on n'hésitera pas à reprendre le point 1, 2 ou 3 si les résultats s'avéraient insatisfaisants.

Formalisation et diffusion :

Les résultats sont formalisés pour être diffusé. Ils ne seront utiles qu'une fois devenus une connaissance partagée. C'est bien là l'aboutissement de la démarche. C'est aussi là que réside la difficulté d'interprétation et de généralisation.

I.4 Processus d'Extraction de Connaissances à partir des Données (ECD) :

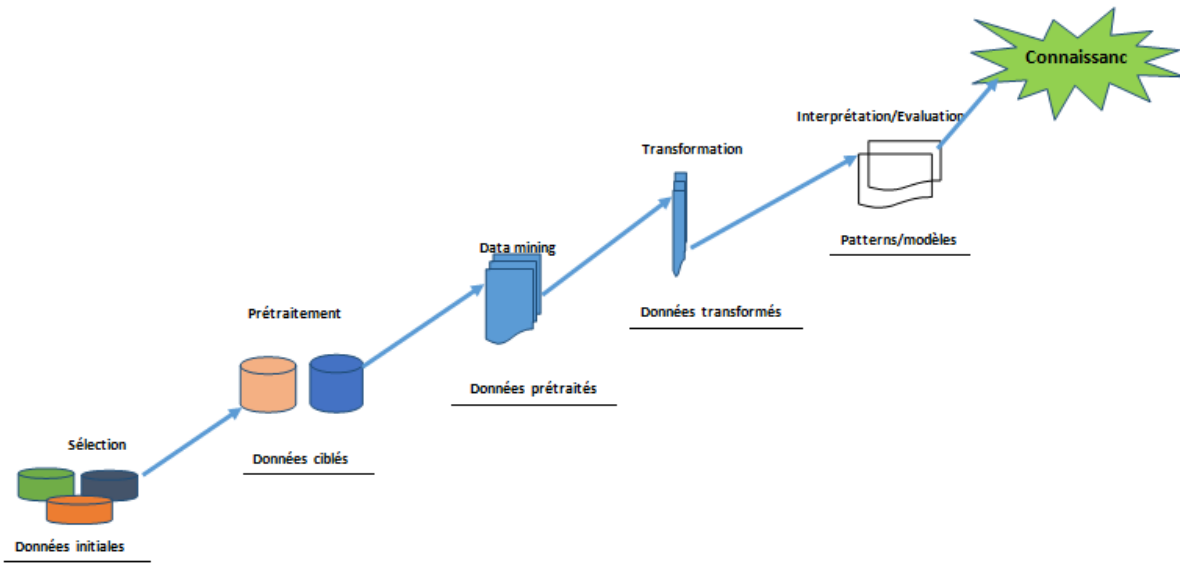

Figure I. 2. Processus d'ECD.

Etapes du processus ECD :

Il existe neuf étapes et sont résumées comme suit :

Compréhension du domaine d'application : consiste à développer une compréhension du domaine d'application et des connaissances pertinentes préalables. Dans cette phase l'analyste prépare pour comprendre et définir les objectifs opérationnels du processus d'ECD.

Création d'un jeu de données ciblées : dans cette phase l'analyste sélectionne les données à utilisées.

Nettoyage des données et prétraitement : dans cette étape on cherche à structurer nos données, en passant par l'élimination des données bruyantes et/ou des valeurs aberrantes, le recueil des informations nécessaires pour modéliser et tenir compte du bruit et choisir les stratégies de traitement des valeurs manquantes.

Réduction et projection des données : c'est pour trouver des attributs utiles pour présenter les données en fonction de l'objectif de la tâche d'extraction, et d'utiliser des méthodes de réduction de dimensionnalité ou de transformation afin de réduire le nombre effectif de variables d'étude et donner de nouvelles variables plus pertinentes.

Choix de la tâche (data mining task) : faire correspondre les objectifs opérationnels du processus d'ECD (étape 1) à une tâche particulière de fouille de données, comme la classification, la régression ou la description et synthèse de données.

Choix de l'algorithme de fouille de données appropriés : c'est pour sélectionner les méthodes à utiliser pour chercher des patterns dans les données, décider quels sont les modèles et paramètres appropriés, et conclure par le choix d'une méthode particulière de fouille de données en accord avec le critère global du processus d'ECD.

Fouille de données (data mining) : il s'agit d'exécuter les méthodes choisies avec leur paramètres afin d'extraire des patterns d'intérêt sous une forme de représentation particulière.

Interprétation des patterns extraits : cette étape comprend l'évaluation et l'interprétation des modèles découverts dans les données. Il peut être nécessaire de retourner à l'une des étapes 1 à 7 pour des itérations éventuelles. Cette étape donne l'occasion de revenir sur les étapes précédentes, mais aussi d'avoir une représentation visuelle des patterns, de supprimer les patterns redondants ou non représentatifs et de transformer le résultat en informations compréhensibles par l'utilisateur final.

Consolidation des connaissances extraites : dans cette phase finale, on utilise les connaissances obtenues et on les intègre dans d'autres systèmes pour des actions ultérieurs, ou bien on les documente et les rapporte aux utilisateurs concernés [4].

I.5 Tâches du data mining :

Pour bien mener un projet, il faut parcourir les taches suivantes :

I.5.1 La classification :

Elle permet de prédire si une instance de données est membre dans un groupe ou d'une classe prédéfinie, la classe étant un champ particulier.

I.5.2 Estimation :

Consiste à estimer la valeur d'un champ à partir des caractéristiques d'un objet. L'estimation peut être utilisée dans un but de classification : il suffit d'attribuer une classe particulière pour un intervalle de valeurs de champ estimé.

I.5.3 La prédiction :

Consiste à estimer un résultat au futur : on cherche à prédire des valeurs futures à partir des valeurs précédentes en se basant sur quelques méthodes.

I.5.4 Les règles d'associations :

Consiste à déterminer les valeurs qui sont associées. Des exemples des taches de règle d'association sont : La grande distribution, la gestion de stocks, le web (pages visitées).

I.5.5 La segmentation :

Consiste à former des groupes (des clusters) en faisant un partitionnement logique de la base de données. Cette tâche est souvent effectuée avant les précédentes pour construire des groupes sur lesquels on applique des tâches de classification ou d'estimation.

I.6 Type des méthodes de Data Mining :

Pour arriver à exploiter les quantités importantes de données, le data mining utilise des méthodes d'apprentissage automatiques. Ces méthodes sont de deux types : les méthodes descriptives et les méthodes prédictives.

I.6.1 Les méthodes descriptives (recherche de patterns) :

Les méthodes descriptives permettent d'organiser, de simplifier et d'aider à comprendre l'information sous-jacente d'un ensemble important de données. Elles permettent de travailler sur un ensemble de données, organisées en instances de variables, dans lequel aucune des variables explicatives des individus n'a d'importance particulière par rapport aux autres. Elles sont utilisées par exemple pour dégager, d'un ensemble d'individus, des groupes homogènes en typologie, pour construire des normes de comportements et donc des déviations par rapport à ces normes telles que la détection de fraudes nouvelles ou inconnues à la carte bancaire ou à l'assurance maladie, pour réaliser de la compression d'informations ou d'images, etc.

Parmi les techniques et algorithmes utilisés dans l'analyse descriptive, on cite :

- Analyse factorielle (ACP et ACM).
- Méthodes des centres mobiles
- Classification hiérarchique
- Classification neuronale (réseau de Kohonen)
- Recherche d'association

I.6.2 Les méthodes prédictives (modélisation) :

La raison d'être des méthodes prédictives est d'expliquer ou de prévoir un ou plusieurs phénomènes observables et effectivement mesurés. Concrètement, elles vont s'intéresser à une ou plusieurs variables définies comme étant les cibles de l'analyse. Par exemple, l'évaluation de la probabilité pour qu'un individu achète un produit plutôt qu'un autre, la probabilité pour qu'il réponde à une opération de marketing direct, celles qu'il contracte une malade particulière et en guérisse, les chances qu'un individu ayant visité une page d'un site web y revienne, sont typiquement des objectifs que peuvent atteindre ces méthodes.

En exploration des données prédictives, il y a deux types d'opération : la discrimination ou classement, et la régression ou prédiction, tout dépend du type de variable à expliquer. La discrimination s'intéresse aux variables qualitatives, tandis que la régression s'intéresse aux variables continues. Les méthodes de classement et de prédiction permettent de séparer des individus en plusieurs classes. Si la classe est connue au préalable et que l'opération de classement consiste à analyser les caractéristiques des individus pour les placer dans une classe, la méthode est dite « supervisée ». Dans le cas contraire, on parle de méthode « non-supervisée ». Ce vocabulaire étant issu de l'apprentissage automatique.

La différence entre les méthodes descriptives de classification que l'on a vues précédemment, et les méthodes prédictives de classement provient du fait que leur objectif est divergent : les premières « réduisent, résument, synthétisent les données » pour donner une vision plus claire en vue de la prédiction des valeurs de ces cibles pour les nouveaux arrivants. Parmi les techniques et algorithmes utilisés dans l'analyse prédictive [5], on cite :

- Arbre de décision
- Réseaux de neurones
- Régression linéaire
- Analyste probabiliste

Les techniques utilisées :

- K-moyennes.
- Apriori.
- K-NN.
- Réseaux de neurones.
- Algorithmes génétiques.
- Chaine de Markov cachées.
- Arbre de décision.
- Réseaux bayésiens.
- Soft computing : ensembles flous.

I.7 La classification

On peut trouver quelques mots en littérature qui portent le même sens de la classification comme la segmentation. Le terme classification est un terme large et très vaste qui peut couvrir plusieurs significations et décrire plusieurs types de problèmes. Plusieurs définitions ont été proposées par les spécialistes. Certains ont un point de vue axé sur l'apprentissage. Ils définissent la classification par « l'action de regrouper en différentes catégories des objets ayant certains points communs ou faisant partie d'un même concept, sans avoir connaissance de la forme ni de la nature des classes au préalable » [6]. Pour d'autres « effectuer une classification, c'est mettre en évidence des relations entre des objets, et entre ces objets et leurs paramètres » [7]. Selon un problème de classification « consiste à affecter des objets, des candidats, des actions potentielles à des catégories ou des classes prédéfinies ». Enfin, « la classification est une méthode d'analyse des données qui vise à regrouper en classes homogènes un ensemble d'observation [8].

I.7.1 Quelques méthodes de la classification

- k-moyennes.
- Apriori.
- K-NN.
- Réseaux de neurones.
- Algorithmes génétiques.
- Réseaux Bayésiens.
- Chaines de Markov cachés.

I.7.2 Les types de classification

Il existe deux types principaux de classification :

Classification supervisée : L'objectif de la classification supervisée est principalement de définir des règles permettant de classer des objets dans des classes à partir de variables qualitatives ou quantitatives caractérisant des objets. Les méthodes s'étendent souvent à des variables Y quantitatives (régression). On dispose au départ d'un échantillon dit d'apprentissage dont le classement est connu. Cet échantillon est utilisé pour l'apprentissage des règles de classement. Il est nécessaire d'étudier la fiabilité de ces règles pour les comparer et les appliquer, évaluer les cas de sous apprentissage ou de sur apprentissage (complexité du modèle). On utilise souvent un deuxième échantillon indépendant, dit de validation ou de test.

Classification non supervisée : Les méthodes de classification non supervisée regroupent les objets en un nombre restreint de classes homogènes et séparées. Homogènes signifie que les éléments d'une classe sont les plus proches possible les uns des autres. Séparées veut dire qu'il y a un maximum d'écart entre les classes. La proximité et l'écart ne sont pas nécessairement au sens de distance. L'homogénéité et la séparation entrent dans le

cadre des principes de cohésion et d'isolation de Cormack [9]. Cet objectif est à distinguer des procédures de discrimination, ou encore déclassement pour lesquelles une typologie est a priori connue, au moins pour un échantillon d'apprentissage.

I.7.3 La classification et la distance entre objets

La classification a pour objectif de regrouper les objets en classes, les objets d'une même classe se ressemblent et les objets de classes différentes ont au contraire peu de points en commun. Toute méthode de classification est ainsi basée sur une mesure de distance ou de similarité-dissimilarité entre objets, et une stratégie d'agrégation qui permettent de construire les classes [10]. De nombreuses méthodes de classification sont disponibles dans les logiciels statistiques : méthodes de partitionnement (K-means, Nuées dynamiques etc.), cartes auto-organisatrices de Kohonen (Self Organizing Maps), méthodes hiérarchiques descendantes et ascendantes….etc. les méthodes de classification ascendante hiérarchique (CAH).

Des centaines de distances ont été proposées pour classer des données, parmi lesquelles la distance euclidienne est la plus populaire. Mais, lorsqu'il s'agit de classer des séries temporelles, l'utilisation de la distance euclidienne, et de toute autre métrique sur les données brutes peut conduire à des résultats peu intuitifs.

I.8 Les séries temporelles

Les séries temporelles, appelées aussi séries chronologiques, sont des collections de mesures ordonnées dans le temps, et constituent une manière structurée pour représenter des données. Donc une série temporelle est une liste de dates, dont chacune est associée à une valeur. Visuellement, il s'agit d'une courbe qui évolue au fil du temps [11]. Par exemple, les ventes quotidiennes d'un produit peuvent être représentées sous forme de série temporelle.

Ce sont des donnés mesurées à des intervalles de temps régulier. On peut les voir comme des suites d'observations répétées d'un même phénomène à des dates différentes (par exemple la température moyenne journalière en un lieu donné, la consommation moyenne en électricité chaque mois, le prix du baril de pétrole chaque jour...). Les dates sont souvent équidistantes (séries mensuelles, trimestrielles ou annuelles) sauf dans quelques cas (cas de données journalières en économie pas toujours disponibles les jours non ouvrables). On représente habituellement une série temporelle (x_t) tel que $\{1<t<T\}$ (t : le numéro de l'observation) à l'aide d'un graphique avec en abscisse les dates et en ordonnée les valeurs observées.

L'étude des séries temporelles correspond à l'analyse statistique d'observation régulièrement espacée dans le temps. Cette étude est appliquée de nos jours dans des domaines aussi variées qu'astronomie, météorologie, médecine ou économie. L'objectif de l'étude de ces séries est de décrire, modéliser, expliquer puis prévoir les phénomènes dans le futur.

Comme indiqué, on représente généralement une série temporelle par des représentations graphiques où les abscisses indiquent le temps et les ordonnées indiquent les valeurs du phénomène étudié. Si la série est stable autour de sa moyenne donc elle est une

série stationnaire, sinon non stationnaire, et quand on aura un phénomène qui se reproduit à des périodes régulières, donc ce phénomène est saisonnier.

I.8.1 Domaines d'application

On trouve des exemples de séries temporelles dans de très nombreux domaines. La liste suivante n'est qu'un échantillon :

Finance et économétrie : évolution des indices boursiers, des prix, des données économiques des entreprises, des ventes et achats de biens, des productions agricoles ou industrielles.

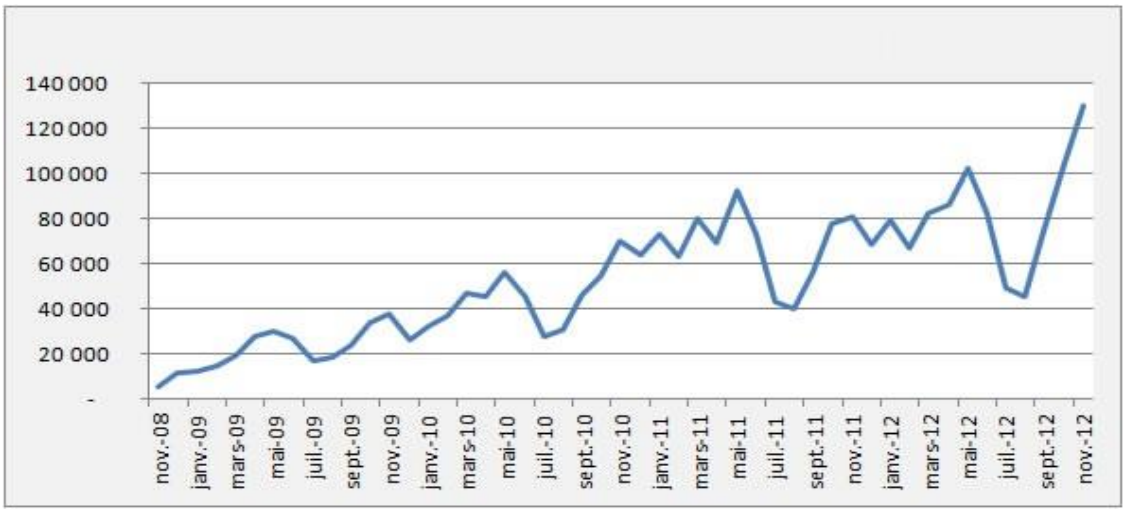

Figure I. 3. Statistiques sur le nombre de pages lues par mois

Assurance : analyse des sinistres.

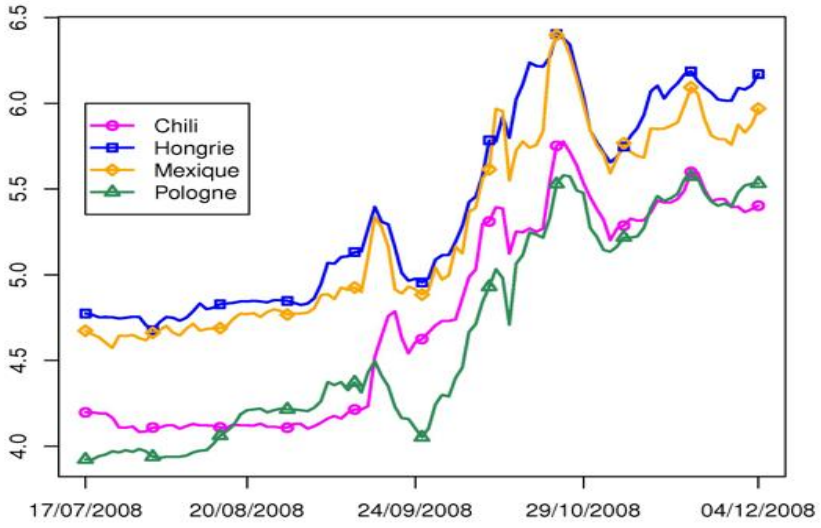

Figure I. 4. Représentation des degrés de séisme au l'échelle de Richter dans 4 pays.

Médecine et Biologie : suivie des évolutions des pathologies, analyse d'électro-encéphalogrammes et d'électrocardiogrammes, suivie de maladies cancéreuses.

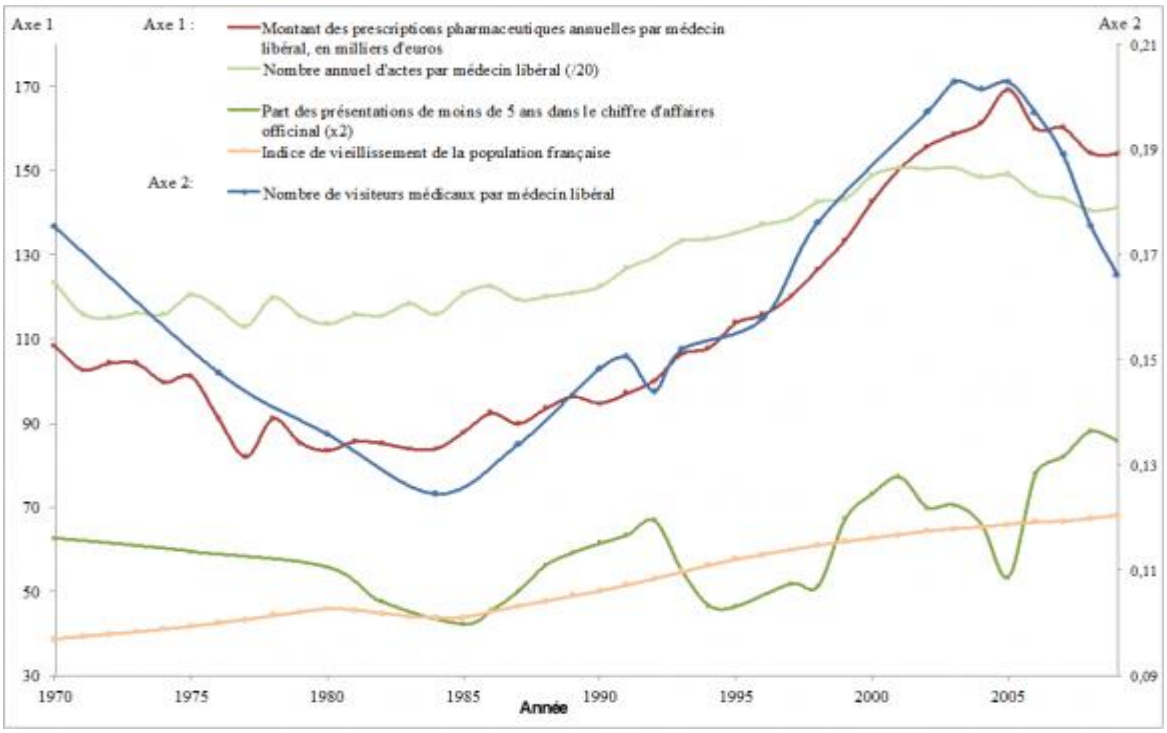

Figure I. 5. Statistique sur 2 axes concernant les médicaments vendus et le nombre de visiteurs chez le médecin libéral.

Science de la terre de l'espace : indices de marées, variations des phénomènes physiques (météorologie), évolution des taches solaires, phénomènes d'avalanches.

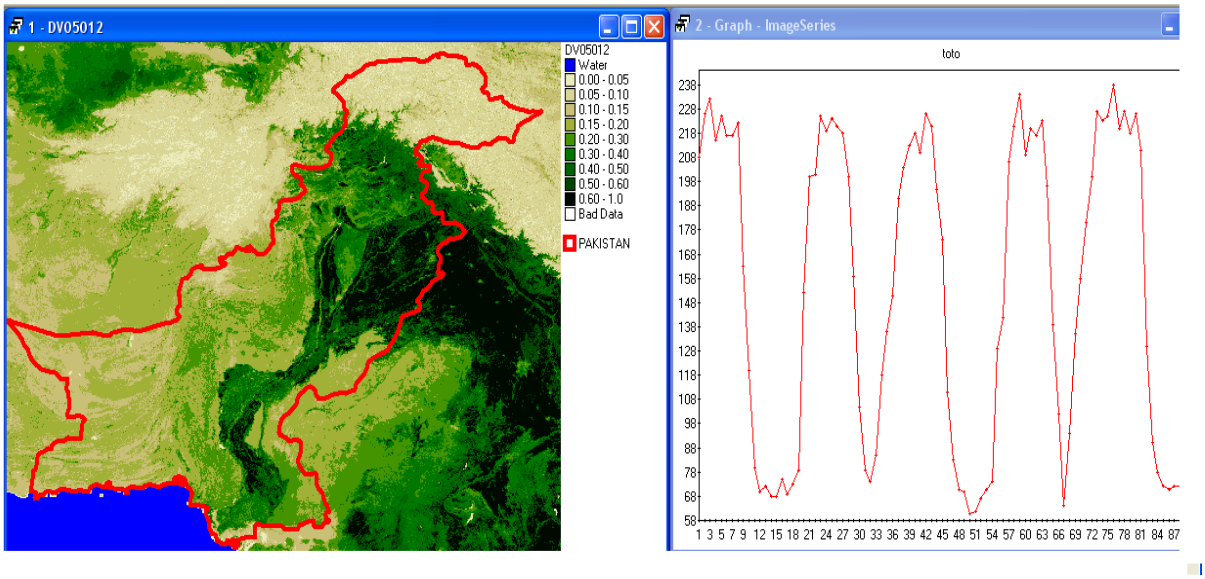

Figure I. 6. Représentation graphique pour une étude géographique (science de la terre).

Traitement de signal : signaux de communications, de radars, sonars, analyse de la parole.

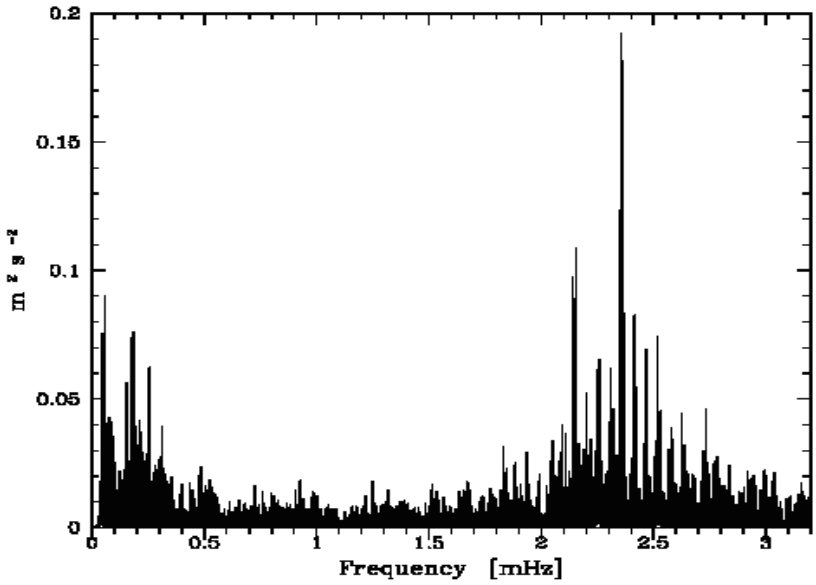

Figure I. 7. Représentation de série de traitement de signal.

Traitement de données : mesures successives de position ou de direction d'un objet mobile (trajectographie).

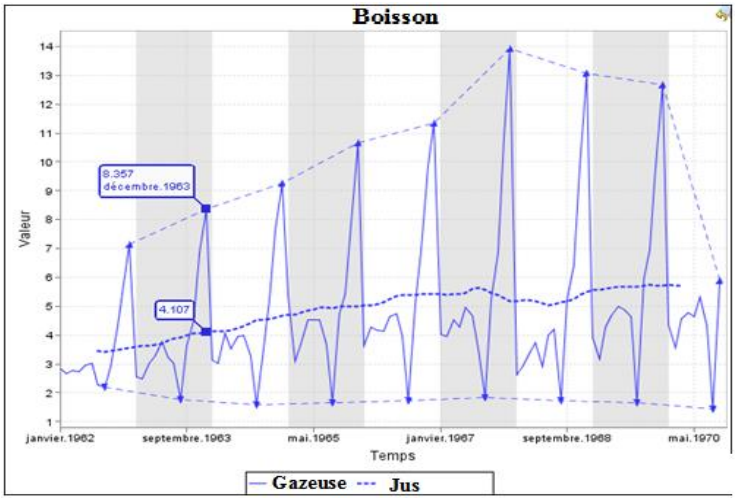

Figure I. 8. Statistique sur les boissons vendues à travers le temps.

Métrologie : variation de phase ou de fréquences des oscillateurs.

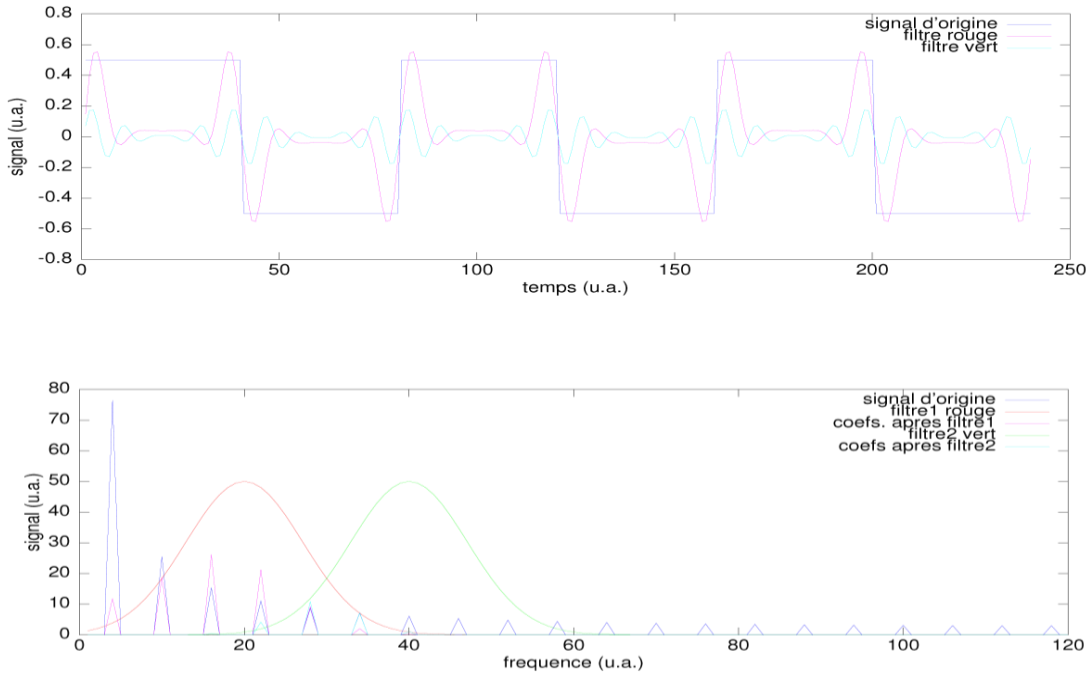

Figure I. 9 Statistique d'une fréquence d'un circuit.

I.9 Conclusion :

Dans ce chapitre, nous avons présenté dans un premier temps l'importance du data mining et son rôle dans différents domaines ainsi que ses objectifs et son emplacement dans le processus d'extraction de connaissances à partir des données. Ensuite, nous avons présenté les différents types de classification, ainsi les méthodes de chaque type. Enfin, nous avons introduit les séries temporelles à travers plusieurs exemples applicatifs. Le prochain chapitre introduira la classification des séries temporelles en particulier.

Chapitre II :Segmentation des séries temporelles

II.1 Introduction :

Un type particulier de la classification est la classification des séries temporelles. Des séquences composées de séries de symboles nominaux d'un alphabet sont appelées généralement des séquences temporelles, et des séquences d'éléments continus à valeurs réelles sont connues comme des séries chronologiques. Les séries temporelles sont notamment classées comme des données dynamiques parce que leurs valeurs caractéristiques changent en fonction du temps, ce qui signifie que les valeurs à chaque point d'une série temporelle sont une ou plusieurs observations qui sont faites par ordre chronologique. Ce sont un type de données qui est naturellement à grande dimension et à grande taille. Les séries temporelles sont d'un intérêt en raison de leur omniprésence dans divers domaines allant de la science, de l'ingénierie, des affaires, la finance, l'économie, la santé, au gouvernement.

Malgré que la série soit composée d'un grand nombre de points de données, elle peut également être considérée comme un objet unique. Le regroupement de tels objets complexes est particulièrement avantageux car il conduit à la découverte de modèles intéressants dans des ensembles de séries de données chronologiques. Comme ces modèles peuvent être soit des motifs fréquents ou rares, plusieurs défis de recherche se sont apparus : le développement de méthodes pour reconnaître les changements dynamiques dans le temps de la série, détection d'anomalie et d'intrusion, contrôle de processus, et la reconnaissance de caractères.

II.2 Classification ou regroupement de séries temporelles :

Pour mettre en évidence l'importance et la nécessité de grouper des ensembles de données de chronologiques, les objectifs de ce regroupement sont donnés comme suit:

- Les bases de données de séries chronologiques contiennent des informations précieuses qui peuvent être obtenues grâce à la découverte de motifs. Le regroupement est une solution commune réalisée pour découvrir ces motifs sur ces ensembles de données chronologiques.
- Les bases de données chronologiques sont très grandes et ne peuvent pas être bien traitées par les humains. Par conséquent, de nombreux utilisateurs préfèrent traiter avec des ensembles de données structurées plutôt que de très grands ensembles de données. En conséquence, les données chronologiques sont représentées comme un ensemble de groupes par agrégation des données en clusters non chevauchant ou par une taxonomie comme une hiérarchie de concepts abstraits.
- Le regroupement des séries chronologiques est l'approche la plus utilisée en tant que technique d'exploration, et aussi comme un sous-programme pour l'algorithmes d'exploration de données les plus complexes, telles que l'indexation, la classification et la détection des anomalies.
- Représenter les structures des séries chronologiques sous forme d'images visuelles (visualisation des données chronologiques) peut aider les utilisateurs à comprendre rapidement la structure des données, les clusters, les anomalies, et d'autres régularités dans des ensembles de données [12].

II.3 Applications de la classification des séries chronologiques :

La classification des séries chronologiques est principalement utilisée pour la découverte de modèles intéressants dans la série elle-même. Elle se divise en deux groupes : le premier groupe est celui qui est utilisé pour trouver des modèles qui apparaissent fréquemment dans l'ensemble de données. Le deuxième groupe contient les méthodes pour découvrir les modèles qui se produisent dans les ensembles de données étonnamment [13].

En bref, trouver les groupes de séries chronologiques peut être avantageux dans différents domaines pour répondre à la suite des problèmes réels suivants: détection d'anomalies, de nouveauté ou de discorde. La détection des anomalies implique des méthodes pour découvrir des modèles inhabituels et inattendus qui se produisent dans des ensembles de données étonnamment. Par exemple, dans les bases de données de capteurs, le regroupement des séries chronologiques qui sont produites par les lectures du capteur d'un robot mobile dans le but de découvrir les événements.

II.3.1 Reconnaissance des changements dynamiques dans la série:

Détection de la corrélation entre les séries chronologiques, par exemple, dans les bases de données financières. Il peut être utilisé pour trouver les entreprises avec le prix d'achat d'actions similaires déménagement.

II.3.2 Prévision et recommandation:

Un regroupement et une fonction d'approximation par cluster peut aider l'utilisateur à prévoir et à recommander. Par exemple, dans les bases de données scientifiques, il peut répondre à des problèmes tels que la recherche des modèles de vent magnétique solaire pour prédire la tendance d'aujourd'hui.

II.3.3 Découverte de Motifs:

Pour découvrir les modèles intéressants dans les bases de données. Par exemple, dans la base de données marketing, différents modèles quotidiens des ventes d'un produit spécifique dans un magasin peuvent être découverts [14].

Le schéma ci-dessous présente quelques applications du regroupement des séries chronologiques dans différents domaines :

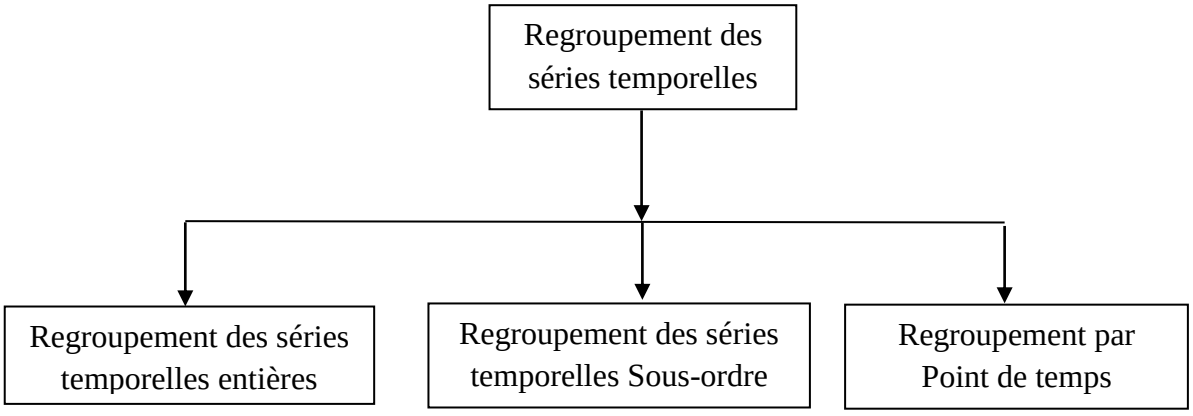

Figure II. 1. Regroupement des séries temporelles

II.4 Les composants de la classification des séries temporelles

Le processus de classification des séries temporelles peut être décomposé comme suit :

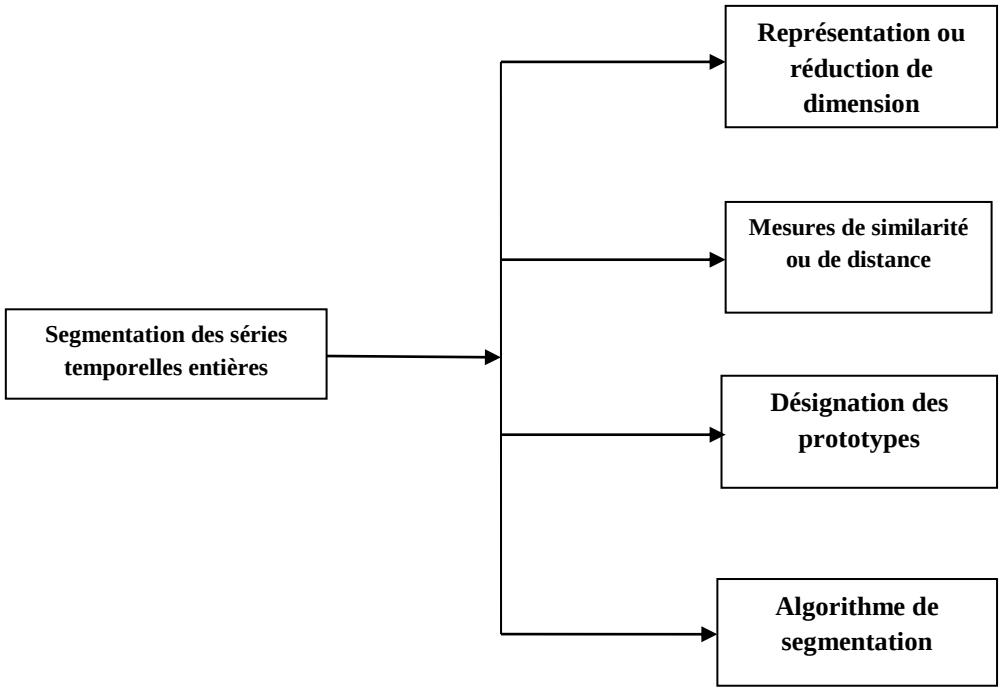

Figure II. 2. Un aperçu des quatre composants de la segmentation des séries temporelles.

II.4.1 Méthodes de représentation

La première composante de la procédure de classification des séries temporelles est la réduction de dimension de la série, qui est une solution commune pour la plus part des approches. Cette composante est connue aussi sous le nom de représentation des séries temporelles. La réduction de dimension représente la série brute (initiale) dans un autre espace en la transformant dans un espace de dimension inférieure ou en remplaçant par une autre plus représentative. La raison que la réduction de dimension est très importante dans la classification des séries temporelles est tout d'abord parce qu'elle réduit les besoins en terme de mémoire car toutes les série brutes ne peuvent pas tenir dans la mémoire principale. Deuxièmement, le calcul de la distance entre les séries brutes est très coûteux en temps de calcul, et la réduction de dimension accélère considérablement la segmentation. Enfin, lorsque l'on mesure la distance entre deux séries temporelles brutes [15], des résultats non intuitifs peuvent être recueillis, car certaines mesures de distance sont très sensibles à certaines « distorsions » dans les données, et par conséquent, en utilisant des séries brutes, on peut regrouper les séries qui sont similaires dans le bruit au lieu de les grouper selon leur similarité en forme. La probabilité d'obtenir un type de cluster différent est la raison pour laquelle le choix de la méthode appropriée pour la réduction de dimension et son rapport est un challenge. En fait, c'est un compromis entre vitesse d'exécution et qualité et tous les efforts doivent être faits pour obtenir une balance de juste équilibre entre temps d'exécution et qualité.

- **Formalisation**

Etant donné une séries temporelles $F_i = \{f_1, ..., f_t, ... f_T\}$, la représentation est la transformation de la série dans un autre vecteur a dimension réduite $F'_i = \{f'_1, ... f'_x\}$ où $x < T$ et si deux séries sont similaires dans l'espace d'origine, alors leurs représentations devraient être similaires dans l'espace de transformation aussi [16].

Choisir une méthode de représentation appropriée peut être considéré comme l'élément clé qui influence l'efficacité et la précision de la solution. La haute dimensionnalité et le bruit caractérisent la plupart des séries de données temporelles, et par conséquent, les méthodes de réduction de dimension sont généralement utilisées afin de faire face à ces deux problèmes et promouvoir la performance. Les techniques de réduction de dimension ont progressé un long chemin et sont largement utilisés pour des séries temporelles à grande échelle et chacune a ses propres avantages et inconvénient. En conséquence, de nombreuses recherches ont été menées en mettant l'accent sur la représentation et la réduction dimensionnalité.

Dans la taxonomie des représentations, il existe généralement quatre types de représentations : adaptatives aux données, non adaptative aux données, basées sur des modèles et imposées par les données. Les quatre approches sont représentées dans le schéma suivant :

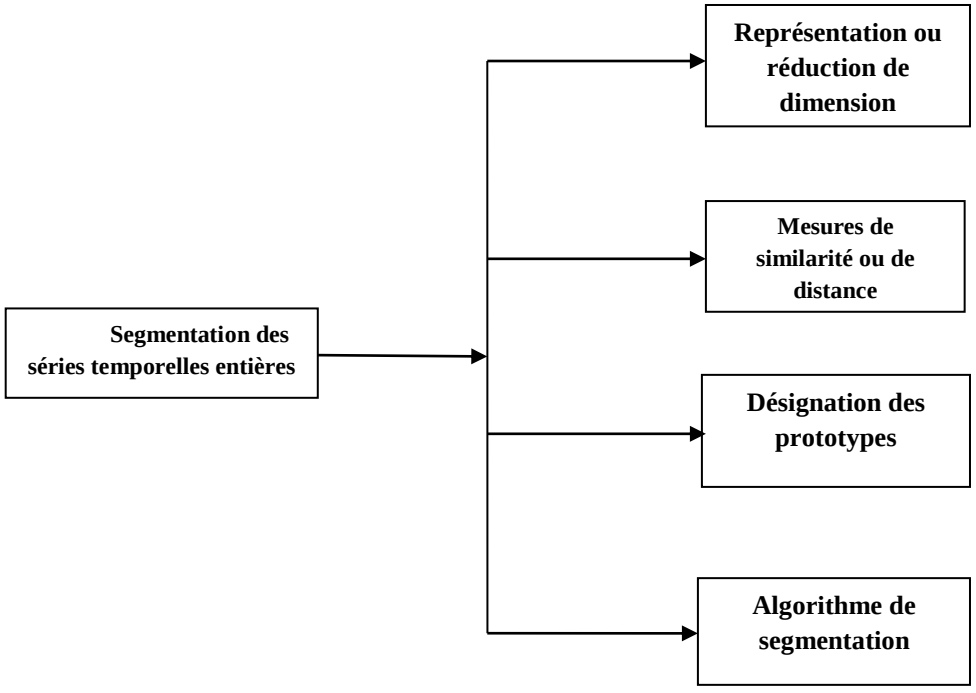

Figure II. 3. Un aperçu des quatre méthodes de représentation des séries temporelles.

II.4.1.1 **Les méthodes adaptatives aux données**

Les méthodes de représentation adaptatives aux donnés opèrent sur toutes les séries de l'ensemble de données et tentent de minimiser l'erreur globale de reconstruction en utilisant une longueur arbitraire pour les segments (non-égaux). Cette technique a été appliquée dans différentes approches telles que l'interpolation polynômiale par morceaux (PPI), la régression polynômiale par morceaux (PPR), approximation linéaire par morceaux (PLA),

approximation constante par morceaux (PCA), approximation constante par morceaux adaptative (APCA), décomposition de valeur singulière (SVD), langage naturel (NL), langage naturel symbolique (NLG), approximation par agrégation symbolique (SAX et iSAX). Les représentations adaptatives aux données peuvent mieux approximer chaque série, mais pour la comparaison de plusieurs séries temporelles c'est plus difficile [17].

II.4.1.2 **Les méthodes non adaptatives aux données**

Les approches non adaptatives aux données sont des représentations qui sont appropriées à la segmentation de séries de taille fixe (de même longueur), et la comparaison entre les représentations de plusieurs séries est simple. Les méthodes dans cette catégorie sont les ondelettes : HAAR, DAUBECHIES, Coeiflets, Symlets, transformé en ondelettes discrète (DWT), polynômes spectrales de Chebyshev, DFT spectrale, Mappages hasardeux, approximation par agrégation par morceaux (PAA) et PLA indexable (IPLA)

II.4.1.3 **Les méthodes basées sur des modèles**

Les approches basées sur des modèles représentent une série d'une manière stochastique tels que les chaînes de Markov et les chaînes de Markov cachées (HMM), modèles statistiques, séries temporelles en Bitmaps, moyenne mobile autorégressive (ARMA). Dans les méthodes adaptative, non- adaptatives aux données et basées sur des modèles, l'utilisateur peut définir le taux de compression en se basant sur le cas d'application en main.

II.4.1.4 **Les méthodes dictées par les données**

En revanche, dans les approches dictées par les données, le taux de compression est défini automatiquement en fonction de séries brutes comme Clipped.

II.4.2 Les mesures de similarité et de di-similarité

La classification des séries temporelles se bas en grande partie sur la mesure de distance. Il existe différentes mesures qui peuvent être appliquées pour mesurer la distance entre des séries temporelles. Certaines sont proposées pour des représentations spécifiques des séries temporelles, par exemple « MINDIST » qui est compatible avec SAX, et certains d'entre elles fonctionnent indépendamment des méthodes de représentation, ou sont compatibles avec les séries de données brutes.

Dans la classification traditionnelle, la distance entre des objets statiques est exactement calculée, mais dans la classification des séries temporelles, la distance est calculée approximativement. En particulier, afin de comparer des séries avec des intervalles d'échantillonnage irréguliers et avec des longueurs différentes, il est d'une grande importance de déterminer la similarité entre les séries. Il existe différentes mesures de distance conçus pour déterminer la similarité entre les séries temporelles. La distance de Hausdorff, la distance de Hausdorff modifiée (MODH), distance basée sur les HMM, Dynamique Time Warping (DTW), la distance euclidienne, la distance euclidienne dans un sous-espace de l'analyse en composante principale PCA, et la plus longue séquence commune (LCSS) [18], sont les méthodes de mesure de distance les plus populaires qui sont utilisés pour les séries temporelles. Une des façons les plus simples pour calculer la distance entre deux séries est de

les considérer comme des séries uni-variées, puis calculer la mesure de la distance à travers tous les points de temps.

- **Définition**

Une série temporelle uni-variée est la forme la plus simple de données temporelles. C'est une suite de nombres réels, représentant des valeurs, recueillies régulièrement dans le temps.

II.4.2.1 **La distance de séries temporelles**

Soit $F_i = \{F_{i1} \ldots\ldots, F_{it} \ldots\ldots, F_{iT}\}$ une série de longueur T. Si la distance entre deux série est définie à travers tous les points de temps, alors $dist(f_i,f_g)$ est la somme de la distance entre les points individuels :

$$dist(f_i,f_g) = \sum dist(f_{iT},f_{gT})$$

Les recherches effectuées sur les mesures de distance basées sur la forme des séries ont généralement à faire à des problèmes tels que le bruit, passage à l'échelle de l'amplitude ou de discontinuités qui sont les propriétés communes des séries temporelles de données. Celles-ci problèmes sont largement étudiés dans la littérature. Le choix d'une approche appropriée pour calculer la distance dépend des caractéristiques des séries temporelles, leurs longueurs, les méthodes de représentation, et bien évidemment sur l'objectif de segmentation [19].

Typiquement, il y a trois objectifs qui, respectivement, nécessitent des approches différentes :

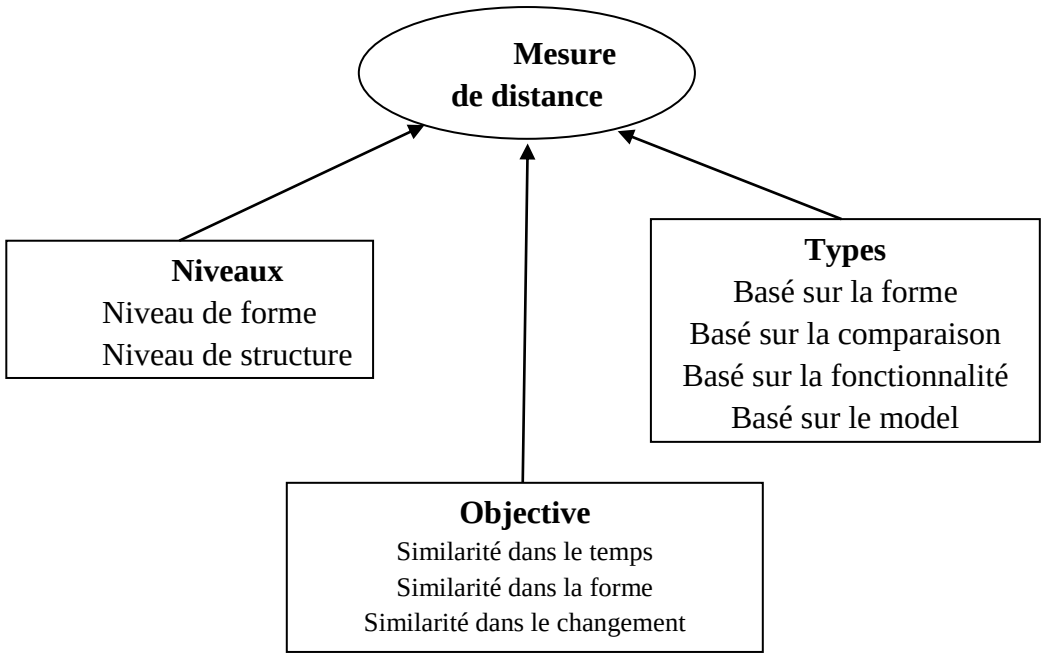

Figure II. 4. Aperçu des mesures de distance.

II.4.2.2 Trouver des séries similaires dans le temps

Parce que cette similarité est sur chaque point du temps, les distances basées sur la corrélation ou la distance euclidienne sont appropriées pour cet objectif. Cependant, parce que ce genre de mesure de distance est coûteux sur les séries brutes, le calcul est effectué sur des séries transformées, comme la transformé de Fourier, les ondelettes ou PAA. La segmentation des séries temporelles qui sont corrélées est classée comme basée sur la similarité dans le temps.

II.4.2.3 Trouver des séries similaires dans la forme

Le temps d'apparition des motifs n'a pas d'importance pour trouver des séries similaire en forme. En conséquence, les méthodes élastiques telles que DTW est utilisée pour le calcul de la dis-similarité. Selon cette définition, des clusters de séries qui ont des motifs similaires de changement sont construit indépendamment des points du temps. La similarité dans le temps est un cas particulier de la similarité dans la forme.

II.4.2.4 Trouver des séries similaires dans le changement (similarité structurelle)

Dans cette approche, généralement des méthodes de modélisation telles que les chaînes de Markov cachées (HMM) ou un processus ARMA sont utilisées, ensuite la similarité est mesurée sur les paramètres du modèle ajusté aux séries temporelles. C'est la segmentation des séries temporelles avec une structure d'auto-corrélation similaire. Cette approche est appropriée pour les longues séries, pas pour des séries modestes ou courtes.

Les approches de segmentation peuvent être classées en deux catégories en fonction de la longueur de la série : niveau de la forme et niveau de la structure. Le niveau de la forme est généralement utilisé pour mesurer la similarité de courtes séries, tandis que le niveau de la structure mesure la similarité qui est basée sur la structure globale ou de haut niveau, et est utilisé pour les longues séries temporelles.

Essentiellement, il existe quatre types de mesure de distance dans la littérature. La mesure de similarité basée sur la forme est pour trouver les séries similaires dans le temps et dans la forme, comme la distance euclidienne, DTW, LCSS, MVM. C'est un groupe de méthodes qui lui sont propres aux courtes séries. La mesure de similarité basée sur la compression est adaptée aux courtes et longues séries, comme CDM, l'auto-corrélation, le coefficient de corrélation de Pearson et les distances connexes. La mesure de similarité basée sur les caractéristiques est approprié pour les longues séries temporelles celle basés sur un modèle est appropriée pour les longues séries temporelles, comme HMM et ARMA.

II.4.3 Les prototypes de segmentation des séries temporelles

La désignation du prototype du cluster ou le représentant du cluster est une procédure essentielle dans les approches de classification des séries temporelles. Une des approches pour aborder le problème de basse qualité des clusters est de remédier à la question de leurs prototypes imprécis, en particulier dans les algorithmes de segmentation tels que k-Means, k-Medoids, C-means floue (FCM), ou même l'algorithme de classification hiérarchique ascendante qui nécessitent un prototype. Dans ces algorithmes, la qualité des clusters dépend fortement de la qualité des prototypes. Etant donné des séries dans un cluster, il est clair que

le prototype du cluster R_j minimise la distance entre toutes les séries temporelles dans le cluster et ce prototype [20].

$$E(C_i,C_j) = 1 / n \sum dist(F_x,R_j),\ C_i=\{F_1,F_2\ldots\ldots,F_n\}$$

Il y a quelques méthodes de calcul des prototypes publiées dans la littérature des séries temporelles, mais la plupart d'entre elles n'ont pas prouvé la justesse de leurs méthodes. Mais, généralement, trois approches peuvent être considérées pour définir les prototypes:

- La séquence médiane de l'ensemble.
- La séquence moyenne de l'ensemble.
- Le prototype de recherche locale.

Ces trois approches sont expliquées et discutées dans ce qui suit.

II.4.3.1 **Utilisation de la médiane comme prototype**

Dans la classification des séries temporelles, la façon la plus courante d'approcher une séquence optimale consiste à utiliser la médiane comme prototype. Dans cette approche, le centre d'un groupe est défini comme étant une séquence qui minimise la somme des carrés des distances aux autres objets se trouvant dans le cluster. Compte tenu des séries dans un cluster, la distance de toutes les paires de séries au sein du cluster est calculée en utilisant une mesure de distance euclidienne ou DTW. Puis, l'une des séries temporelles dans le cluster, qui a la somme de l'erreur quadratique inférieure est définie comme médiane du cluster. De plus, si la distance utilise une approche non élastique telle que la distance euclidienne, ou si le centre de gravité du cluster peut être calculé, on peut dire que la médiane est la série temporelle la plus proche au centre. La médiane du cluster est très fréquente dans les publications liées à la classification des séries temporelles et a été utilisée dans de nombreux articles.

II.4.3.2 **Utilisation de prototype de moyennage**

Si les séries temporelles sont de longueur égale et la métrique de distance est non-élastique (par exemple, la distance euclidienne) alors, la méthode de calcul de moyenne est une simple technique de moyennage qui est égale à la moyenne de la série à chaque point. Cependant, dans le cas où il existe des séries temporelles de longueur différente ou dans le cas où la similitude entre les séries de temps est basée sur une similitude de forme, la nature du mappage une à une, le laissera incapable de capturer la forme moyenne actuelle. Par exemple, dans les cas où DTW ou LCSS sont très appropriées, le prototype de moyennage est écarté, car ce n'est pas une tâche triviale. Grand nombre de travaux dans la littérature évitent l'utilisation d'approches élastiques (par exemple, DTW ou LCSS) où il est nécessaire d'utiliser un prototype sans donner de motifs suffisants (la classification étant basée sur la similarité dans le temps ou dans la forme) [21].

II.4.3.3 **Utilisation de prototype de recherche local**

Dans cette approche, dans un premier temps la médiane dans le cluster est calculée, puis en utilisant la méthode de calcul de la moyenne, le prototype de moyennage est calculé.

II.4.4 Algorithmes de classification

Dans cette section, les travaux existants liés à la classification des données de série temporelle sont discutés. Certains d'entre eux utilisent les séries dans leur forme brute et certains essaient d'utiliser les méthodes de réduction avant la segmentation. Comme il est montré dans la Figure ci-dessous, la classification peut être classée en générale en six groupes :

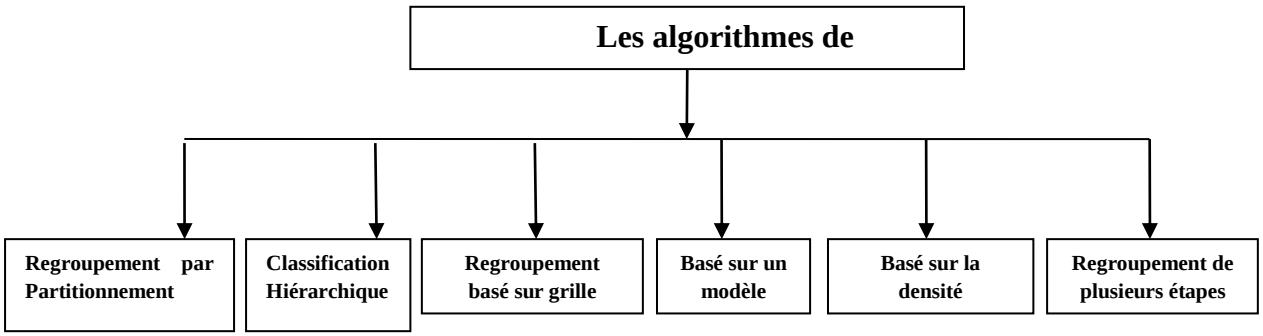

Figure II. 5. Les approches de classification.

II.4.4.1 **Regroupement par partitionnement**

Une méthode de classification par partitionnement fait *k* groupes depuis *n* objets non marqués de manière à ce que chaque groupe contienne au moins un objet. L'un des algorithmes les plus utilisés dans le regroupement par partitionnement est k-Means où chaque groupe a un prototype qui est la valeur moyenne de ses objets. La principale idée derrière le regroupement k-Means est la minimisation de la distance totale (généralement la distance euclidienne) entre tous les objets dans un cluster de leur prototype.

II.4.4.2 **La classification hiérarchique**

La classification hiérarchique est une approche qui construit une hiérarchie de clusters utilisant des algorithmes agglomératifs ou de division. Les algorithmes agglomératifs considère chaque élément comme un cluster, puis fusionne progressivement les clusters (bottom-up) [22]. En revanche, les algorithmes de division commencent avec tous les objets dans un seul cluster, puis divise le cluster pour atteindre des clusters avec un objet (top-down). En général, les algorithmes hiérarchiques sont faibles en termes de qualité, car ils ne peuvent pas ajuster les clusters après le fractionnement d'un cluster dans les méthodes de division, ou après la fusion dans les méthodes agglomératives. En conséquence, les algorithmes de classification hiérarchiques sont généralement combinés avec un autre algorithme dans une approche de regroupement hybride afin de remédier à ce problème. Par ailleurs, certains travaux étendus sont effectués pour améliorer les performances de classification hiérarchique [23].

II.4.4.3 **Basé sur les modèles**

Le regroupement basé sur les modèles tente de récupérer le modèle d'origine à partir d'un ensemble de données. Cette approche suppose un modèle pour chaque cluster, et trouve le meilleur ajustement des données à ce modèle. En détail, elle suppose qu'il ya des centres de gravité choisis au hasard, puis un peu de bruit est ajouté à eux avec une distribution normale. Le modèle que l'on récupère à partir des données générées définit les clusters [24].

Typiquement, les méthodes basées sur des modèles utilisent soit des approches statistiques, comme les réseaux de neurones.

II.4.4.4 **Regroupement basé sur la densité**

Dans le regroupement à base de densité, les clusters sont des sous-espaces d'objets denses qui sont séparés par des sous-espaces, dans lesquels les objets ont une faible densité. L'un des algorithmes connus qui fonctionne par le concept basé sur la densité est DBSCAN où un cluster est étendu si ses voisins sont denses [25]. OPTICS est un autre algorithme basé sur la densité qui aborde la question de la détection de clusters significatifs dans les données.

II.4.4.5 **Regroupement basé sur grille**

Les méthodes basées sur grille quantifient l'espace en un nombre fini de cellules qui forment une grille, puis effectuent le regroupement sur les cellules de la grille. STING et Wave Cluster sont deux exemples typiques d'algorithmes de classification qui sont basées sur le concept basé sur grille [26].

II.4.4.6 **Regroupement en plusieurs étapes**

Bien qu'il existe de nombreuses études pour améliorer la qualité des approches de représentation, de mesure de distance, et prototypes, quelques travaux mettent l'accent sur l'amélioration des algorithmes et présentent un nouveau modèle (généralement comme une méthode hybride) pour le regroupement des données de séries temporelles.

II.4.5 La méthode SAX (Symbolic Aggregate approXimation):

SAX est une représentation symbolique pour les séries chronologiques qui permet la réduction de la dimensionnalité et l'indexation avec une mesure de distance inférieure. Dans les tâches d'exploration de données classiques telles que le clustering, la classification supervisée, l'indexation, etc., SAX est aussi une bonnes représentation par rapport à d'autres aussi bien connues telles que Discrete Wavelet Transform (DWT) et Transformation de Fourier discrète (DFT), tout en nécessitant moins d'espace de stockage. En outre, cette représentation permet aux chercheurs de profiter de la richesse des structures de données et algorithmes en bioinformatique ou text mining, et fournit également des solutions à de nombreux défis associés aux tâches actuelles d'exploration de données. Un exemple est le découverte de motif, un problème que nous avons défini pour les données de séries chronologiques. Il y a un grand potentiel pour l'extension et l'application de la représentation discrète sur une large classe de tâches d'exploration de données [19].

II.4.5.1 **Principe de SAX**

Dans un objectif de réduction de la dimensionnalité, il est important de définir des unités temporelles permettant de regrouper les points des séries temporelles. On définit généralement des intervalles du domaine de définition temporel des séries temporelles à représenter. Très généralement, en raison du coût minime d'acquisition et de stockage des données, les séries temporelles sont enregistrées dans les bases de données sous la forme la plus détaillée possible, indépendamment de l'échelle de temps à laquelle se développent les comportements à identifier. On pourra alors regrouper les points en intervalles sans perdre d'information essentielle. C'est le principe de la représentation symbolique SAX.

SAX est une représentation symbolique de séries univariées centrées réduites qui n'est pas adaptative :

- Le domaine temporel est divisé en intervalles de même taille
- Les classes d'équivalence des valeurs prises par les séries temporelles sont fixées a priori en fonction du nombre de symboles à utiliser, de façon à obtenir un découpage en classes de même effectif sous réserve que la distribution centrée et réduite des valeurs soit normale.

Le non adaptativité de la représentation et la référence à la distance euclidienne donne à SAX les avantages suivants :

- La construction des représentations est extrêmement efficace en temps de calcul ($O(N)$) pour une représentation d'une série temporelle de N points.
- Toutes les représentations (basées sur un même nombre de symboles et des intervalles de même taille) sont trivialement commensurables.

En revanche, cette représentation souffre d'inconvénients qui sont intrinsèquement liés à :

- L'erreur de modélisation qui, pour une réduction donnée de la dimensionnalité, n'est pas minimale puisque le modèle n'est pas localement adapté aux données
- Les classes d'équivalence auxquelles les symboles sont associés ne sont pas forcément pertinentes car elles ne sont pas adaptées aux données [20].

II.4.6 La distance DTW (Dynamic Time Warping):

La mesure de distance Dynamic Time Warping (DTW) est une technique qui est connue depuis longtemps dans la communauté de reconnaissance vocale. Elle permet une cartographie non linéaire d'un signal à un autre en minimisant la distance entre les deux. DTW a été introduite dans la communauté de l'exploration de données comme un utilitaire pour diverses tâches pour les problèmes de séries chronologiques, y compris la classification. La technique a prospéré, en particulier au cours des dernières années, et a été appliquée à une variété de problèmes dans diverses disciplines [21].

DTW est un algorithme d'alignement de séries temporelles développé à l'origine pour la reconnaissance de la parole. Il vise à aligner deux séquences de vecteurs caractéristiques en déformant l'axe du temps itérative jusqu'à une correspondance optimale (selon une métrique appropriés) entre les deux séquences est trouvé.

Les deux séquences peuvent être disposées sur les côtés d'une grille, avec une sur le dessus et l'autre sur le côté gauche. Les deux séquences commencent en bas à gauche de la grille :

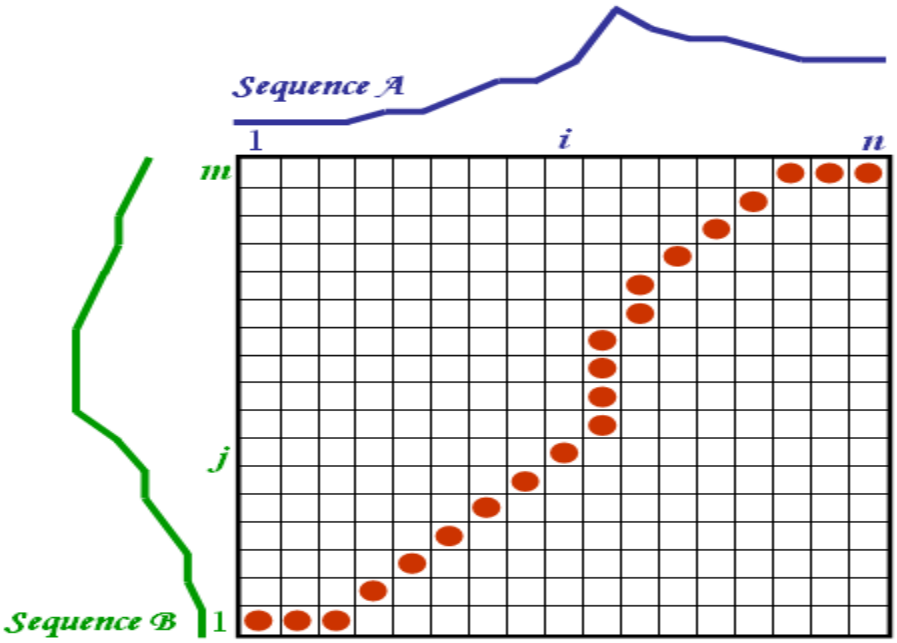

Figure II. 6. La grille dans DTW.

A l'intérieur de chaque cellule, une mesure de distance peut être placée, en comparant les éléments correspondants des deux séquences. Pour trouver le meilleur match ou alignement entre ces deux séquences, on a besoin de trouver un chemin à travers la grille qui minimise la distance totale entre eux. La procédure de calcul de cette distance globale consiste à trouver tous les chemins possibles à travers la grille et pour chacun calculer la distance totale. La distance d'ensemble est le minimum de la somme des distances entre les différents éléments sur le chemin divisée par la somme de la fonction de pondération. La fonction de pondération est utilisée pour normaliser la longueur du trajet. Il est évident que pour toutes les séquences considérables longues le nombre de chemins possibles à travers la grille sera très grand. Les optimisations majeures ou les contraintes de l'algorithme DTW découlent des observations sur la nature des chemins acceptables par la grille [22].

II.5 Conclusion :

Dans ce chapitre, nous nous sommes focalisé particulièrement sur la classification des séries temporelles, connues aussi sous le nom de regroupement ou clustering. Cette tâche a été décomposée en quatre composante, et nous avons détaillé chacune d'elles à part. Dans le chapitre suivant, nous allons entamer les détails de conception et d'implémentation informatique de notre outil de classification.

Chapitre III : Conception et Implémentation

III.1 Introduction

Pour atteindre l'objectif de classification des séries temporelles, nous avons choisis à chaque composante de cette classification une méthode et nous l'avons implémenté. Ainsi, notre choix a été porté dans la première composante sur la méthode SAX, et dans la deuxième sur une variante de la distance euclidienne. Les détails de ces choix feront l'objet de ce chapitre.

III.2 Représentation en fichiers CSV :

Tout d'abord, nous notons que nos données seront prises de fichiers CSV. Un fichier CSV est un simple fichier texte dans lequel les valeurs sont séparées par une virgule, ce qui permet de sauvegarder les données dans un format de tableur. Chaque ligne comporte le même nombre de valeur (des champs) et parfois ces valeurs sont entourées de guillemets anglais (" ").

À la place des virgules, les données peuvent également être séparées par des points-virgules, par une tabulation ou par le symbole « | » et, là encore, peuvent être entre guillemets. Ce qui donne par exemple :

"Prénom", "Nom", "Email", "Age", "Jean", "Petit", "jean@monsite.fr", "34".

Il n'existe pas de structure standard pour les fichiers CSV. En revanche, la majorité des logiciels et systèmes utilisent des règles similaires [23].

III.3 La normalisation des valeurs des séries temporelles

La normalisation permet de rendre une variable statistique centrée réduite. En probabilités et statistiques, une variable centrée réduite est une variable aléatoire dont on a modifié les valeurs afin de fixer sa moyenne et sa variance.

- Centrer une variable consiste à soustraire son espérance à chacune de ses valeurs initiales, soit retrancher à chaque donnée la moyenne (c'est ce qui s'appelle un centrage). Elle constitue simplement en un changement d'origine, qui place la moyenne de la distribution au point 0 de l'axe des abscisses1.
- Réduire une variable consiste à diviser toutes ses valeurs par son écart type

Elle passe par les étapes suivantes :

- On calcule la moyenne générale pour chaque série et l'écart type qui se présentent sous les formules suivantes :

La formule de la moyenne :

$$\bar{\mathbf{x}} = \frac{\mathbf{x1 + x2 + \ldots + xn}}{\mathbf{n}} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{x_i})$$

Avec :

$\bar{X}$: La moyenne

$X_1, X_2 \ldots \ldots X_n$: Les valeurs de la série.

N : le nombre des valeurs de la série.

La formule de l'écarte type :

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Avec :

σ **:** Le symbole de l'écart type.

X_i *:* Les valeurs de la série.

- Ensuite, chaque valeur de la série est substituée de la moyenne et divisée par l'écart type.

La formule finale de la normalisation :

$$x' = \frac{xi - \bar{x}}{\sigma}$$

Une nouvelle série est alors constituée, dont les valeurs se trouvent dans l'intervalle [-1, +1].

III.4 Approche d'approximation :

Les séries temporelles constituent un domaine très actif de la fouille de données. En effet, les bases de données de séries temporelles sont caractérisées non seulement par leur très grand volume, mais aussi par le fait que les informations recherchées ne sont pas directement accessibles à partir des données brutes. Pour cette raison, des changements de représentation doivent être effectués. Nous nous intéressons plus particulièrement à des représentations symboliques, plutôt que numériques, car elles sont intelligibles par les utilisateurs. Nous présentons tout d'abord un cadre général permettant de formuler une très large gamme de représentations symboliques, notamment SAX (Symbolic aggregate approximation) qui est une représentation symbolique de séries temporelles classiquement utilisée.

SAX a été développée pour réduire la dimensionnalité d'une série temporelle en une courte chaine de caractère. SAX suit un processus de deux étapes :

- Piecewise Aggregate Approximation (PAA)
- La conversion d'une séquence de PAA dans une série de lettres.

III.4.1 Algorithme PAA (Piecewise Aggregate Approximation) :

Réduction de la dimension :

On note une séries temporelle $X = x_1, \ldots\ldots x_n$ et l'ensemble des séries temporelles qui constituent la base de donnée $Y = \{y_1, \ldots\ldots y_k\}$.

Sans perte de généralité, on suppose que chaque séquence dans *Y* est de longueur de *n* unités.

Soit *N* la dimension de l'espace transformé.

Une série temporelle *X* de longueur *n* est représenté dans l'espace *N* par un vecteur :

$$X' = x'_1, \ldots\ldots, x'_n$$

Par exemple, une série temporelle composée de huit (n) points est projeté en deux dimensions (N). La série est divisée en deux (N) trames et la moyenne de chaque trame est calculée. Un vecteur de ces moyens devient les données de la représentation réduite.

Les spécialistes de ce domaine ont proposé indépendamment PAA, où, dans chaque séquence les données sont divisées en *k* segments à longueur égale et la valeur moyenne de chaque segment est utilisée en tant que coordonnées d'un vecteur caractéristique à *k* dimensions.

Les avantages de cette transformation sont que :

- Elle est très rapide et facile à mettre en œuvre.
- L'indice peut la construire dans le temps est linéaire.

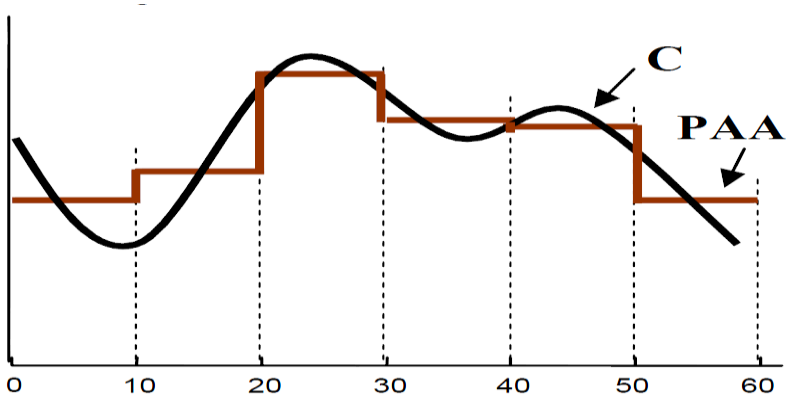

Figure III. 1. Une série de temporelle C représenté par PAA.

III.4.2 Représentation symbolique SAX :

Dans un objectif de réduction de la dimensionnalité, il est important de définir des unités temporelles permettant de regrouper les points des séries temporelles. Le tableau suivant contient les points de coupures des intervalles de valeurs obtenues par PAA et à représenter par le même symbole dans SAX :

								0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**0**
1	0,43	0,67	0,84	0,97	1,07	1,15	1,22	,28	1,34	1,38	1,43	1,47	1,5	1,53	1,56	1,59	1,62	1,64
2	,43		0,25	0,43	0,57	0,67	0,76	0,84	0,91	0,97	1,02	1,07	1,11	1,15	1,19	1,22	1,25	1,28
3		,67	,25		0,18	0,32	0,43	0,52	0,6	0,67	0,74	0,79	0,84	0,89	0,93	0,97	1	1,04
4			,84	,43	,18		0,14	0,25	0,35	0,43	0,5	0,57	0,62	0,67	0,72	0,76	0,8	0,84
5				,97	,57	,32	,14		0,11	0,21	0,29	0,37	0,43	0,49	0,54	0,59	0,63	0,67
6					,07	,67	,43	,25	,11		0,1	0,18	0,25	0,32	0,38	0,43	0,48	0,52
7						,15	,76	,52	,35	,21	,1		0,08	0,16	0,22	0,28	0,34	0,39
8							,22	,84	,6	,43	,29	,18	,08		0,07	0,14	0,2	0,25
9								,28	,91	,67	,5	,37	,25	,16	,07		0,07	0,13
10									,34	,97	,74	,57	,43	,32	,22	,14	,07	
11										,38	,02	,79	,62	,49	,38	,28	,2	,13
12											,43	,07	,84	,67	,54	,43	,34	,25
13												,47	,11	,89	,72	,59	,48	,39
14													,5	,15	,93	,76	,63	,52
15														,53	,19	,97	,8	,67
16															,56	,22		,84
17																,59	,25	,04
18																	,62	,28
19																		,64

Tableau III. 1. Les points de coupures des intervalles

Si l'on dispose des ressources nécessaires, il peut donc être intéressant de construire des représentations symboliques adaptives. De nombreuses représentations symboliques peuvent être envisagées, non seulement en fonction des données à analyser, mais aussi en fonction des tâches d'analyses à effectuer.

III.4.3 Calcul de la distance :

La prochaine phase de segmentation est de calculer la distance entre les séries pour mesurer la similarité. Et dans ce vaste domaine, il existe beaucoup de méthodes mais seulement pour le calcul numérique. Dans notre travail nous nous intéressons à la distance entre symboles, ce qui nous pousse à utiliser une table prédéfinie dont la suivante est un bref exemple :

	A	**B**	**C**	**D**
A	0	0.25	0.75	0.5
B	0.25	0	0.65	0.45
C	0.75	0.65	0	0.21
D	0.5	0.45	0.25	0

Tableau III. 2. Exemple de table de distances entre quatre symboles.

III.4.4 Cas d'application :

Notre cas d'application concerne la surveillance des paramètres réels d'un équipement, ce qui constitue une application de reconnaissance de séquences temporelles réelles. Dans ce type d'application, l'évolution d'un signal capteur représente une succession de paramètres de type réel, et un palier de dégradation peut représenter une séquence réelle bien particulière qui sera mémorisé sous formes de prototypes. Le résultat d'une telle reconnaissance sera le déclanchement d'une pré-alarme, ce qui permet à un expert de localiser la dégradation et prendre ainsi des décisions pour des actions préventives.

III.4.4.1 **Les moteurs d'avions à turbine à gaz**

Les moteurs d'avions sont des exemples courants de systèmes pouvant générer un ensemble suffisant de données pertinentes décrivant leur historique de fonctionnement surtout avant une défaillance. Les données utilisées pour l'expérimentation sont générées par la simulation de fonctionnement de plusieurs moteurs, la provocation de plusieurs cas de défaillances, et la collection des données représentants l'état de chaque système avant de tomber en panne. Cette base de données a été utilisée comme base du challenge lors de la compétition sur les données du pronostic au PHM'08[1].

La figure ci-dessous présente le diagramme simplifié d'un des moteurs avec ses composants essentiels :

[1] Elle est téléchargeable à l'adresse : http://ti.arc.nasa.gov/tech/dash/pcoe/prognostic-data-repository

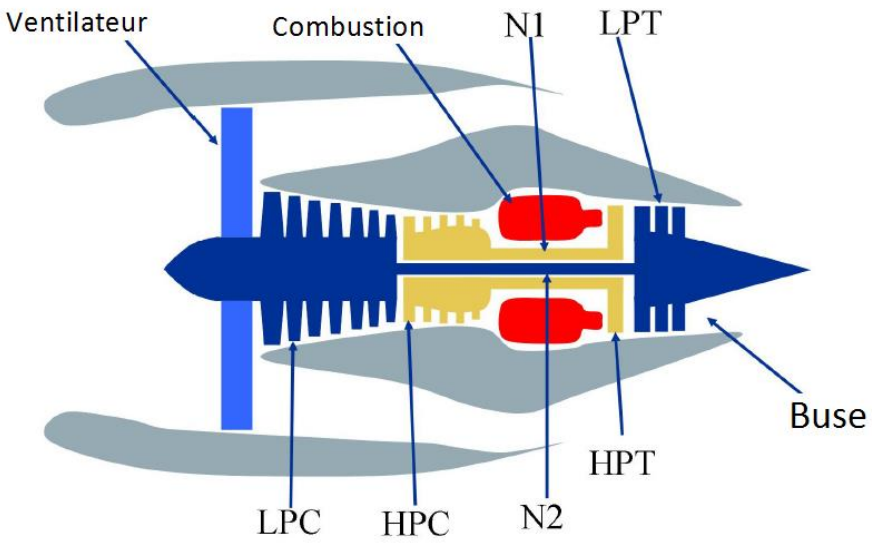

Figure III. 2. Schéma simplifié du moteur à turbine à gaz.

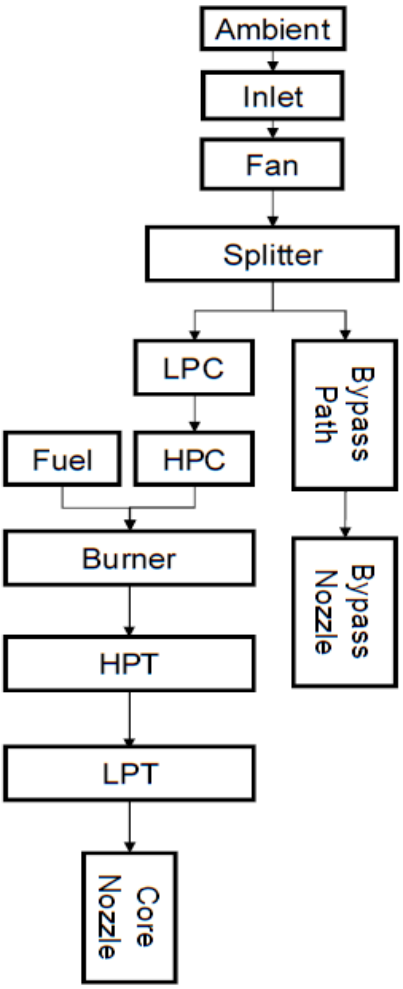

Figure III. 3. Une mise en page montrant les différents modules et de leurs connexions comme modélisée dans la simulation.

Le système utilisé pour la simulation accepte plusieurs entrées permettant de simuler les effets des défaillances et des dégradations à n'importe quel des cinq composants rotatifs du moteur : Ventilateur (Fan), LPC (Low Pressure Compressor), HPC (High Pressure Compressor), HPT (High Pressure Turbine), LPC (Low Pressure Turbine). Parmi ces composants, c'est le HPC qui est ciblé pour subir une dégradation puis une défaillance. Les sorties du système incluent plusieurs réponses des capteurs, dont 21 ont été retenues dans la base de données. Le tableau suivant récapitule ces variables :

Description	Unité	Symbole
HPC efficiency modifier		HPC_eff_mod
HPC flow modifier		HPC_flow_mod
HPCpressure-ratio modifier		HPC_PR_mod
Total temperature at fan inlet	°R	T2
Total temperature at LPC outlet	°R	T24
Total temperature at HPC outlet	°R	T30
Total temperature at LPT outlet	°R	T50
Pressure at fan inlet	Psia	P2
Total pressure in bypass-duct	Psia	P15
Total pressure at HPC outlet	Psia	P30
Physical fan speed	Rpm	Nf
Physical core speed	Rpm	Nc
Engine pressure ratio (P50/P2)		epr
Static pressure at HPC outlet	Psia	Ps30
Ratio of fuel flow to Ps30	pps/psi	phi
Corrected fan speed	Rpm	NRf
Corrected core speed	Rpm	NRc
Bypass Ratio		BPR
Burner fuel-air ratio		farB
Bleed Enthalpy		htBleed
Demanded fan speed	Rpm	Nf_dmd
Demanded corrected fan speed	Rpm	PCNfR_dmd
HPT coolant bleed	lbm/s	W31
LPT coolant bleed	lbm/s	W32

Tableau III. 3. Les différentes variables de la simulation.

III.4.4.2 **La préparation des données**

La base de données contient quatre ensembles de données pour différentes conditions de vol, et divers composants ayant subi une défaillance. Parmi ces ensembles, nous avons choisi le premier qui contient des données sur 100 unités fonctionnant au niveau de mer avec des défaillances concernant uniquement le module HPC. Cet ensemble consiste en multiples séries temporelles multi variées issues de moteurs différents (une flotte de moteurs de même type). Chaque moteur commence en fonctionnement normal au début de chaque série, puis développe un défaut à un point donné durant la série. Le défaut croit en ampleur jusqu'à la défaillance du système. En plus des 24 variables citées au dessus, l'ensemble contient le numéro de l'unité, et le temps (en cycles) de la prise de chaque mesure. A ces variables, on a ajouté une colonne représentant la différence entre le temps de la prise de chaque mesure et celui de la dernière en chaque série. Ensuite, on a procédé à l'étiquetage des données. Pour ce fait, on s'est référé à la visualisation des paramètres pour choisir les exemples positifs et négatifs. Après cette étape, on a supprimé les cinq premières variables de l'ensemble. La première représente seulement un numéro incrémental de l'unité en cours, tandis que la deuxième représente le numéro de cycle, qui a été remplacé par la période citée au dessus. Les trois variables qui suivent représentent les entrées utilisées pour la simulation.

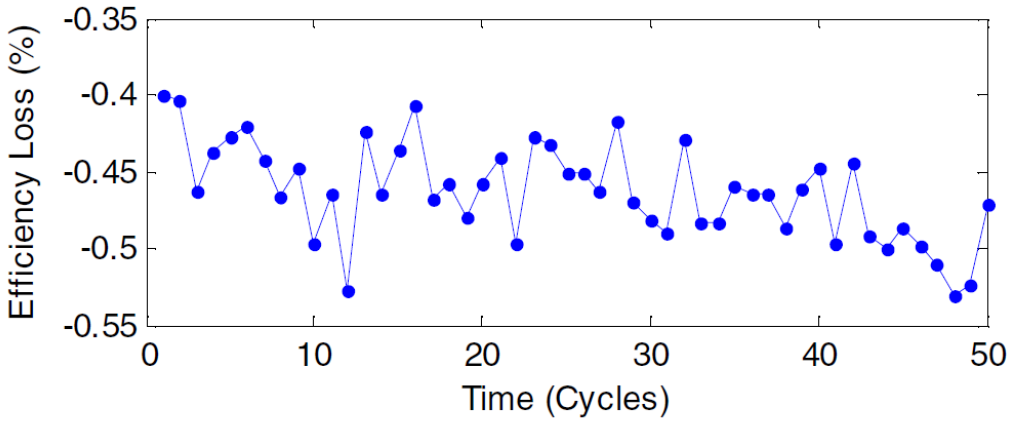

Figure III. 4. Changement dans le paramètre de l'efficacité.

Parmi les variables caractéristiques de ce moteur (vitesse, pression, écoulement, efficacité…etc.), nous allons étudier la pression, représentée dans 100 séries temporelles. On prend un extrait des données et on applique les différentes étapes que nous avons vu précédemment :

1	554.36	**16**	553.94
2	553.75	**17**	553.80
3	554.26	**18**	553.20
4	554.45	**19**	554.18
5	554.00	**20**	554.81
6	554.67	**21**	554.08
7	554.34	**22**	553.63
8	553.85	**23**	553.98
9	553.69	**24**	553.49
10	553.59	**25**	554.00
11	554.54	**26**	554.11
12	554.52	**27**	554.07
13	553.44	**28**	554.68
14	553.48	**29**	554.25
15	554.64	**30**	554.37

Tableau III. 4 Echantillon représentant une série temporelle.

Figure III. 5. Les séries temporelles dans notre outil.

En ce moment, on rend cette série centrée réduite en appliquant les différentes étapes de la normalisation (calcul de la moyenne et de l'écart type…) :

La moyenne :

$$\bar{X} = \frac{554.36+553.75+554.26+\cdots\ldots\ldots\ldots\ldots\ldots\ldots+554.25+554.37}{30} = \frac{17176.28}{30} = 572.54.$$

L'écart type :

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})}^{2}$$

$$\sigma = \sqrt{\frac{1}{30}}(554.36\text{-}572.54) \dots\dots\dots\dots\dots\dots\dots\dots(554.37\text{-}572.54) = \sqrt{341.223}.$$

$$\text{donc } \sqrt{341.223} = 18.472.$$

Après, on doit remplacer chaque valeur (échantillon) par la valeur normalisée, ce qui veut dire qu'on doit appliquer sur chaque valeur la formule :

$$x' = \frac{xi - \bar{x}}{\sigma}.$$

Par exemple : $x' = \frac{554.36 - 572.54}{18.472}$ = -0.984.

Donc la séries deviennent normalisées, le tableau suivant présente la série normalisé :

1	-0.984	**16**	-0.906
2	-0.917	**17**	-0.014
3	-0.989	**18**	-0.046
4	-0.979	**19**	-0.993
5	-0.903	**20**	-0.959
6	-0.967	**21**	-0.999
7	-0.985	**22**	-0.023
8	-0.911	**23**	-0.904
9	-0.920	**24**	-0.931
10	-0.925	**25**	-0.903
11	-0.974	**26**	-0.997
12	-0.975	**27**	-0.999
13	-0.933	**28**	-0.966
14	-0.931	**29**	-0.990
15	-0.960	**30**	-0.983

Tableau III. 5 La série après normalisation.

Figure III. 6. La normalisation dans notre outil.

Voici une représentation graphique de notre outil des séries normalisées :

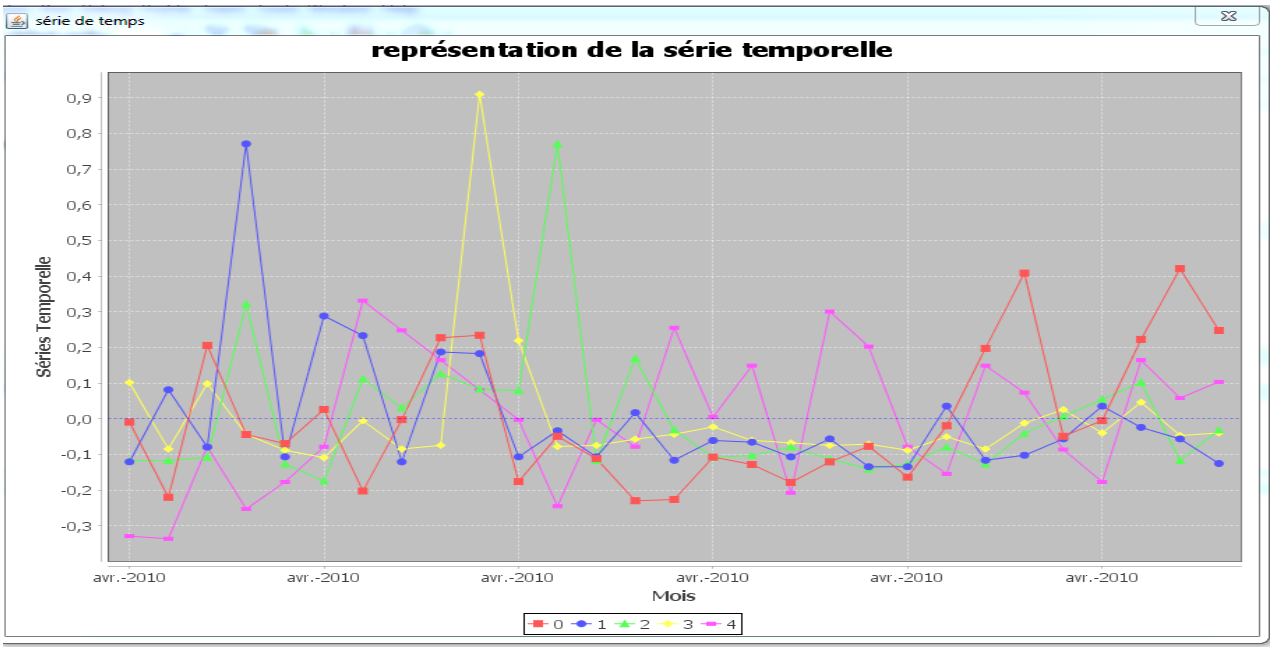

Figure III. 7. La visualisation des séries normalisées.

Ensuite, on applique l'algorithme PAA avec un $k = 6$, c.à.d. qu'on va deviser notre série qui contiens 30 valeurs par 5 segments dont chacun contient 6 valeurs. Après on calcule pour chaque segment sa moyenne, et on le remplace par celle-ci :

Exemple : $\bar{X} = \frac{-0.984+(-1.017)+(-0.989)+(-0.979)+(-1.003)+(-0.967)}{6} = \frac{-5.939}{6}$ = -0.9893.

On continue ce calcul jusqu'à que la série devienne de taille réduite :

1	-0.989
2	-0.998
3	-0.915
4	-0.901
5	-0.989

Tableau III. 6. La série après PAA.

La dimension de la série qui contient 30 valeurs a été diminuée jusqu'à 5 valeurs, ce qui nous montre l'intérêt essentielle de l'approche PAA.

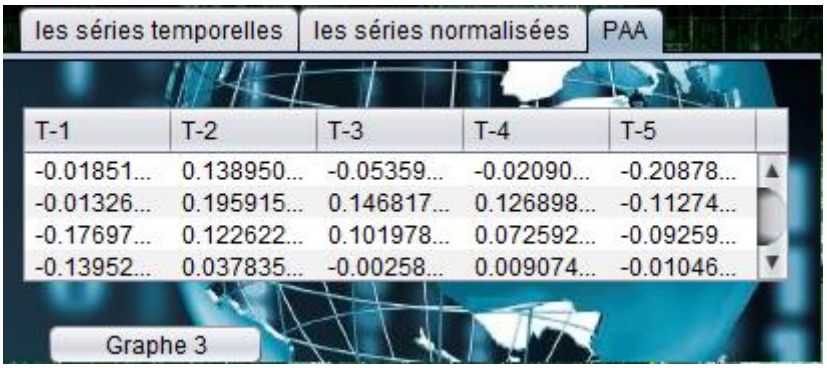

Figure III. 8. Les résultats de l'algorithme PAA.

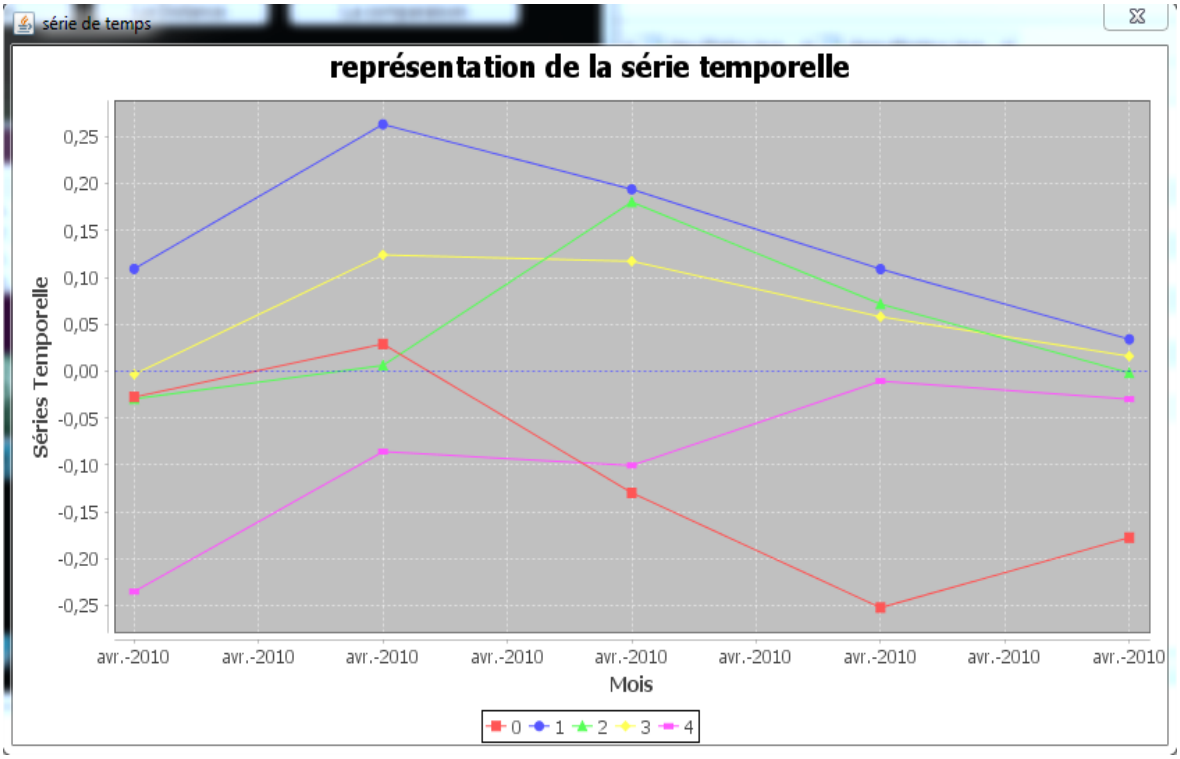

Figure III. 9. Les résultats de PAA sous forme graphique.

Passons maintenant à la différence entre PAA et SAX qui est de convertir nos résultats numériques qu'on a obtenus par l'approche PAA aux symboles comme montré précédemment (chaque intervalle de moyenne est remplacé par un symbole).

Notre série après l'algorithme de SAX :

1	A
2	B
3	B
4	D
5	C

Tableau III. 7. La série après SAX.

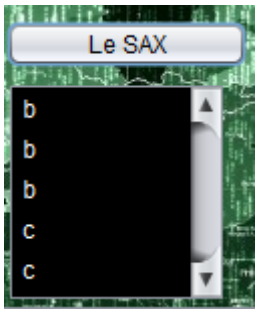

Figure III. 10. Les symboles après SAX.

En procédant de la même manière avec toutes les séries de la base, nous allons nous retrouvés avec de séries symboliques, ce qui nous permettra d'appliquer l'algorithme de segmentation pour classer ces séries. Mais d'abord, on doit au premier lieu calculer les distance entre toutes les séries après avoir fixé un seuil de distance pour que nous puissions faire les comparaisons et en prend juste les séries dont la distance est inférieure à ce seuil.

La distance utilisée est une variante de la distance euclidienne :

Figure III. 11. Les résultats du calcul de la distance.

Maintenant après avoir fixé le seuil (dans cet exemple on a fixé le seuil à 0.7) et faire les comparaisons on obtient :

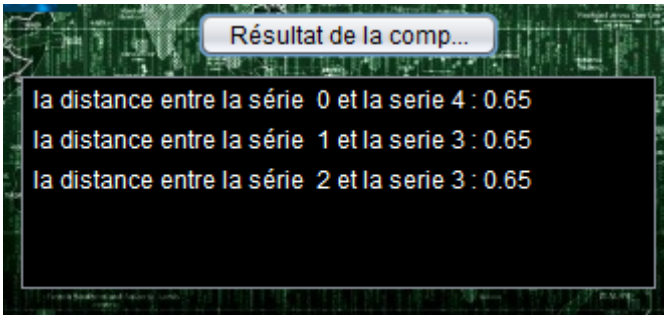

Figure III. 12.Les résultats de la comparaison.

III.5 Conclusion :

Une fois les séries temporelles représentées « symboliquement » et la distance calculée, un algorithme de classification peut être utilisé pour obtenir les classes des séries, et ensuite un prototype représentatif de chaque classe peut être choisit. En revenant à la forme initiale des séries, la similarité entre les séries d'une même classe peut être clairement visualisée. Mais une similarité identifiée avec mois de ressources en termes de calcul.

Conclusion Générale

La classification est considérée comme l'une des tâches accomplisse dans un processus d'extraction de connaissances à partir de données, qui est connu souvent sous le terme de data mining ou de fouille de données en français. La classification étant l'action de regrouper en différentes catégories des objets ayant certains points communs ou faisant partie d'un même concept, sans avoir connaissance de la forme ni de la nature des classes au préalable. Il existe beaucoup de méthodes et d'approches pour procéder à la classification d'objets de différentes natures, et chaque approche dépends de la nature des objets sur lesquels elle procède.

Dans ce travail, nous nous sommes attaqués au problème de la classification des séries temporelles, des collections de mesures ordonnées dans le temps, constituant une manière structurée pour représenter des données. La particularité de ce type de données réside dans leurs grands volumes, ce qui constitue un réel obstacle pour l'application des méthodes classiques de classification : ces méthodes ne sont pas adaptées à prendre en compte de telles quantités de données. C'est pour ce fait, que nous nous sommes basés sur une décomposition du processus de classification en plusieurs composantes pour arriver à notre fin.

Dans un premier lieu, nous avons procédé à la représentation des séries dans un autre espace, plus petit que celui d'origine. Ceci permet de réduire la dimensionnalité des données, et travailler sur des volumes de données raisonnables. Parmi les avantages de cette approche, on trouve la minimisation des ressources utilisées en termes de vitesse de calcul et espace de stockage. Néanmoins, et vu la complexité de cette première phase, nous nous sommes contenté de la réaliser uniquement, avec une approche plus simple de la phase qui la suit, le calcul de la distance, puis la classification.

Ceci nous ouvre la perspective de continuer notre travail, en étudiant avec plus de détails les autres composantes de la classification, en complétant notre outil par des implémentations qui complètent sa tâche qu'il prévoit atteindre : la classification des séries temporelles.