République Algérienne Démocratique et Populaire
Ministère De L'enseignement Supérieur Et De La Recherche Scientifique
Université Amar Télidji Laghouat

Faculté des sciences et sciences de l'ingénierie

Département d'informatique

Projet de Fin d'Etudes

En vue de l'obtention d'un diplôme de master en informatique
Option : systémes d'information et de décision

Thème :

vers un calcul de distance entre le comportements des objets dans les techniques de classification

Présenté par : Maroua Lahreche

Composition du jury :

Mr. LAKHDARI Abdellah	Président	UATL
Mr. BOUAKAZ Mustapha	Examinateur	UATL
Mr. BENARFA Abdelmadjid	Examinateur	UATL
Mr. Guellouma Younes	Encadreur	UATL

juin 2015

Remérciement

je tiens tout d'abord a remercier Dieu le tout puissant et miséricordieux, qui m'a donné la force et la patience d'accomplir ce modeste travail.

Je tiens à remercier particulièrement mon encadrant Mr. Younes Guellouma d'avoir accepté d'encadrer ma recherche ; Vos précieux conseils, vos explications et orientations qui m'ont éclairé la méthodologie de la recherche et votre aide durant toute la période de la réalisation de ce travail.

Mes vifs remerciements vont également aux membres du jury pour l'acceptation d'examiner mon travail.

Je profite l'occasion pour exprimer mes sincères remerciements, généralement, à tous mes enseignants qui ont veillé à me donner du savoir tout au long de ma carrière universitaire

Dédicace

Je dédie ce travail

A Mes très chers parents ;

Vous êtes perpétuellement à mes cotés pour me soutenir et m'encourager.

Aucun hommage ne saurait exprimer la grandeur de mon amour, mon estime, et ma profonde reconnaissance pour les sacrifices et les efforts consentis pour moi. J'espère avoir été a la hauteur de vos attentes, et que Dieu vous garde afin que je puisse vous combler à mon tour.

A mes fréres et mes belles soeurs ;

Que ce modeste travail soit le net reflet de ma reconnaissance pour le soutien moral, l'affection fraternelle, ainsi que la compréhension et les encouragements dont vous avez toujours fait preuve. Que Dieu voue protège et guide vos pas avec mes voeux de succès et bonheur à l'éternité.

A toute ma famille ;

Je me permets, par le biais de ce modeste travail, de vous apporter la chaleur de notre affection, témoignage de liens affectueux qui m'attache.

A tous mes amis ;

Vous m'avez offert ce qu'il y a de plus cher : la sincère amitié.

Abstract

The calculation of the distance between two objects i,in a very useful mean in various domains such as statics,DATA MINING ,TEXT MINING , etc...

However,in certain situations this calculation doesn't provide the desired results due to many parameters(factors) such as size,representation,coding,noise(effects), etc...
the aim of this document is to study and compare two very used techniques in the changes of the shape of the objects whilst keeping their representation.

the goal to attain first is to apply a comparison between the regression and P.C.A in the context of the change of the representation and then an implementation of a case study is to be planned(envisaged.

Résumé

Le calcul de distance entre les objets est un moyen très utile dans divers domaines tel que les statistiques, le data mining, le texte mining ...etc mais dans certaines situations, ce calcul ne donne pas des rèsultats acceptables a cause de beaucoup de problémes comme la taille, la représentativité, le codage, les bruits.

Le but de ce document est d'etudier et de comparer deux techniques trés utilisées dans le changement de la forme des objets tout en gardant leur représentativité. L'objectif à atteindre est d'abord d'effectuer une comparaison entre la régression et l'ACP dans le contexte du changement de la représentation, ensuite une implémentation menée d'une étude de cas est envisagée.

resume arabe

Table des matières

Table des figures

Chapitre 1

Intrduction générale

le Data Mining a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques.
Text Mining C'est un ensemble de traitements informatiques consistant à extraire des connaissances selon un critère de nouveauté ou de similarité dans des textes produits par des humains pour des humains. Dans la pratique, cela revient à mettre en algorithmes un modele simplifi des théories linguistiques dans des systèmes informatiques d'apprentissage et de statistiques.
Web Mining est l'application des techniques d'exploration de données en vue de découvrir des constantes, schémas ou modèles, dans les ressources d'internet ou les données le concernant. Selon ses cibles, la fouille du web peut être divisée en trois types : la fouille de l'usage du web, la fouille du contenu du web, la fouille de la structure du web.
Le problème fondamental posé par ces trois c'est le calcul de la similarité on utilise comme solution Les distances classiques (pour les espaces vectoriels) Distance de Manhattan, Distance euclidienne (pour les textes)

Problématique
Le problème du calcul de similarité est du à :

1. **Qualitatif Vs Quantitatif**Les données sont géralement en qualité et non pas en quantité. Pour résoudre ce problème on doit transformer les données qualitatives en données quantitatives.
2. **Non linéarité**Les données sont non linéaires, pour cela on doit les rendre linéaire en utilisant la technique des noyaux
3. **Taille d'objet variable**Comme la taille de l'objet est variable on utilise la technique de régression

4. **Codage d'objet**Les objets ne sont pas nécessairement en nombre c'est-à-dire ce sont des ensembles, graphes, arbres, ou semi anneaux. Pour le codage on utilise la technique des noyaux rationnels.

Champs d'étude On s'intéresse au problème ou les données ne sont pas sur la mme longueur
Le grand défi est de calculer une bonne distance qui représente le mieux la similarité entre ses objets en tenant compte de

1. La représentativité des objets
2. La sensibilité envers les données éloignées

Des techniques statistiques permettant d'extraire une forme de comportement "behaviour" comme la moyenne $\overline{X}$ et l'ecart type σ variance $V(x)$ et des techniques plus complexes comme :

- **Régression : la méthode des moindres carrés** s'agit d'ajuster un nuage de points $M_i(x_i, y_i)$ $i = 1, 2, 3, ..., n$ selon une relation linéaire. La méthode des moindres carrés consiste à minimiser la somme des carrés des écarts, écarts pondérés dans le cas multidimensionnel, entre chaque point du nuage de régression et son projeté, parallélement a l'axe des ordonnées, sur la droite de régression.
- **ACP : L'Analyse en composantes principales** est une méthode de la famille de l'analyse des données et plus généralement de la statistique multivariée qui consiste a transformer des variables corrélées (liées) en nouvelles variables décorrélées les unes des autres. Ces nouvelles variables sont nommées "composantes principales". Elle permet de reduire le nombre de variables et de rendre l'information moins redondante.
- **AFD : L'analyse factorielle discriminante** est une technique statistique qui vise a décrire, expliquer et prédire l'appartenance à des groupes prédéfinis d'un ensemble d'observations a partir d'une série de variables prédictives.

En informatique, pour la reconnaissance optique de caracteres. L'analyse discriminante est utilisée pour reconnaitre un caractére imprimé a partir d'informations simples, comme la présence ou non de symétrie, le nombre d'extrémités...
Dans ce mémoire, on va discuter deux techniques, la regression et leACP et on doit faire une comparaison entre ces deux techniques en utilisant des données de tailles différentes.

Chapitre 2

Généralité sur le calcul de similarité et le data mining

2.1 Introduction

Dans ce chapitre nous allons prsenter les notions de base du Data Mining ainsi que les algorithmes utilisant la similarité et nous allons parler brièvement sur les distances classiques, la segmentation en prenant le K-means comme exemple et puis la classification en parlant sur les SVM et plus précisément les noyaux.
Apres nous allons définir les problèmes liées aux ces trois derniers et leurs solutions.

2.2 Inrtoduction au Data Mining

Le Data Mining est un nouveau champ situé au croisement de la statistique et des technologies de l'information (bases de données, intelligence artificielle, apprentissage etc.) dont le but de découvrir des structures dans de vastes ensembles de données.[14]
L'objectif du Data Mining est particulièrement important, on peut le voir comme la question à laquelle on souhaite répondre à partir des données.

L'étape de préparation des données est également essentielle, notamment du fait de l'hétérogénéité des données (on peut à la fois travailler sur des données structurées, comme les bases des données relationnelles, et sur des données non structurées, comme du son ou de la vidéo par exemple). Le but de cette étape est d'organiser et de classer les données en vue de les utiliser lors de l'étape suivante (on estime cette étape à environ 40% de la charge de travail d'un projet de Data Mining). Vient ensuite l'étape d'élaboration et de choix des modèles à appliquer (modèles issues de l'intelligence Artificielle, des statistiques...etc.) sur les données, en vue d'en extraire les connaissances recherchées dans le cadre de l'objectif initial. Enfin, le Data Mining nécessite le plus souvent l'intervention d'un expert métier, pour évaluer, contrôler et exploiter les connaissances extraites. En effet, puisque ces connaissances sont le résultat de traitements semi-automatiques voire automatiques, il est nécessaire de les valider, ce qui ne peut se faire sans comprendre le sens des données en entrée.[22]

le Data Mining est une exploration d'une grande quantité de données (centaines de variables/milliers d'observations) en vue de rechercher des modèles relationnels entre des variables et ensuite de valider ces modèles en les appliquant sur de nouvelles données.

Dans les méthodes utilisées par le Data Mining, on distingue deux grandes familles d'algorithmes :

Les méthodes descriptives permettent d'organiser, de simplifier et d'aider à comprendre l'information à partir des sources de données. (Par exemple : recherche d'associations / recherche de séquences similaires ...etc)

Les méthodes prédictives visent à expliquer ou prévoir plusieurs phénomènes observables et effectivement mesurés. On cherche à prédire la valeur d'une variable cible à partir des valeurs de prédicateurs[9]. (Par exemple : régression linéaire multiple / réseaux de neurones / arbres de régression...)

Autrement dit on cherche à anticiper la valeur de quelque chose (par exemple, si un client risque de ne pas pouvoir rembourser un prêt, c'est la variable cible) en fonction de ses caractéristiques connues (âge, emploi, salaire... ce sont les prédicateurs), en se basant pour cela sur les données dont on dispose (les précédents clients et les valeurs des prédicateurs et des variables cibles) [17]

Dans le Data Mining plusieurs techniques utilisent la notion de « similarité » entre les entités du modèle étudié. Dans la partie suivante on donne une définition à cette notion importante et quelques algorithmes.

2.3 Algorithmes utilisant la similarité

Il en existe plusieurs algorithmes, dont :

Le regroupement hiérarchique : chaque objet est regroupé dans un même cluster avec l'ensemble d'objets les plus similaires à lui,cette approche est incrementale et peut se faire en amont(chaque cluster contient un seul objet)Ou en aval (tous les objets sont dans le même cluster).

l'algorithme EM : L'algorithme espérance-maximisation est une classe d'algorithmes qui permettent de trouver le maximum de vraisemblance des paramètres de modèles probabilistes lorsque le modèle dépend de variables latentes non observables.[1]

l'analyse en composantes principales : permet d'analyser les objets en les projetant sur un espace de dimension minimal.

K-means : l'algorithme des k-means consiste à regrouper dans des cluster tous les objets similaires.

Knn : permet de determiner la nature d'un objet en se réf''erant aux objets similaires déjà connus.

...

Notion de similarité

La représentation des individus dans les domaines variés nécessitant l'apprentissage automatique (entrepôts de données) n'est pas toujours sous forme quantitative, or que les algorithmes d'apprentissages étudiés ont besoin de ce genre de représentation pour assurer le calcul des distances, pour le faire, chaque individu doit être représenté quantitativement par l'intermédiaire de vecteurs d'attributs, de matrices ou de formes multidimensionnelles.

On dit que deux objets X et Y décrits par attributs sont similaire s'il y a un nombre d'attributs qui ont identiques entre X et Y qui vérifient :[4]. Bien que la vérification de la similarité entre deux objet ne semble pas une tâche facile, le moyen le plus efficace de calculer ce degré et de définir une mesure de distance entre les objets en allant d'une représentation quantitative des attributs. On peut exprimer la distance entre deux objets x et y en fonction du degré de similarité [5],[18] :

$$Dist(x, y) = 1 - Similarite$$

2.3.1 Les distances classiques

On peut définir de plusieurs manières la distance entre deux points, bien qu'elle soit généralement donnée par la distance euclidienne (ou 2-distance).

Soient $X(x_1, x_2, ..., x_n)$ et $Y(y_1, y_2, ..., y_n)$ les représentations de deux objets avec n attributs quantitatives, il existe une longue liste de fonctions distance applicables selon le cas[6] :

- **Distance de Manhattan (appelée aussi Hamming)**

$$Dist(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

- **La distance euclidienne**

$$Dist(x, y) = \sqrt{\sum_{i=1}^{n} sqrt(x_i - y_i)}$$

- **La distance de Minkowski**

$$Dist(x, y) = p\sqrt{\sum_{i=1}^{n} (x_i - y_i)^p}, p > 0$$

– **Distance de Tchebychev**

$$Dist(x,y) = Max_{i=1.....n}|x_i - y_i|$$

– **Séparation angulaire**

$$Dist(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}$$

– **Distance de Canbirra**

$$Dist(x,y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{x_i + y_i}$$

– **Distance de Sebestyen**

$$d^2(x_1,x_2) = (x_1 - x_2)^t W (x_1 - x_2)$$

(W =matrice diagonale de pondération)

– **distance de Mahalanobis**

$$d^2(x_1,x_2) = (x_1 - x_2)^t C^{-1} (x_1 - x_2)$$

(C = matrice de variance-covariance)

nous avons cités les différentes distances classiques en donnant leurs formules, et par la suite nous présentons quelques techniques basées sur la distance dans la classification supervisée et non supervisée.

2.3.2 la segmentation"classification non supervisée"

Consiste â former des groupes homogènes à l'intérieur d'une population (l'ensemble des objets sur lequel on fait l'expérience où chaque objet représente un individu) homogène, contrairement è la classification les groupes ne sont pas préétablis.
Pour cette tâche il n'y a pas de valeur à estimer ou à prédire mais il s'agit de maximiser l'homogénéité à l'intérieur de chaque groupe et la minimiser entre les groupes c.-à-d. maximiser l'hétérogénéité entre les groupes.[26] Pour résoudre certains problèmes complexes, il peut s'avérer utile de commencer par segmenter la population (la diviser en groupes) en espérant que le problème soit alors plus simple à résoudre sur les groupes ainsi constitués. La segmentation est une tâche d'apprentissage "non supervisée" car on ne dispose d'aucune autre information préalable que la description des exemples.

Plusieurs tâches sont utilisées dans la segmentation comme : k-means, Algorithmes hiérarchiques, Réseaux de neurones. Nous prenons ici la méthode des k-means car elle est très simple à mettre en œuvre et très utilisée. Elle comporte de nombreuses variantes et elle est souvent utilisée en combinaison avec d'autres algorithmes.

K-means Appelé aussi C moyenne ou méthode des centres mobiles, le K-means est un algorithme de clustering non supervisé largement utilisé dans le Data Mining et a fait ses preuves dans plusieurs domaines (Bases de données, marketing, entrepôts de données ...). Cet algorithme consiste à regrouper les informations en K classes selon leurs similarités, par conséquent utilise la notion de distance.
L'algorithme du K-means s'écrit de la façon suivante [19],[12] :

1. Choisir K centres initiaux ;
2. Répartir chaque individu dans le groupe i dont ci est le centre le plus proche ;
3. Si aucun élément ne change de groupe alors arrêter le calcul ;
4. Calculer les nouveaux centres et aller à 2.

Si K est le nombre de clusters, I le nombre d'itérations,Sle nombre de composantes représentant chaque individu ($S = n(resp.nXm)$ si l'individu est représenté par un vecteur (resp. une matrice)) et T le nombre d'individus, la complexité du calcul des distances est de $\mathcal{O}(K \times I \times S \times T)$.

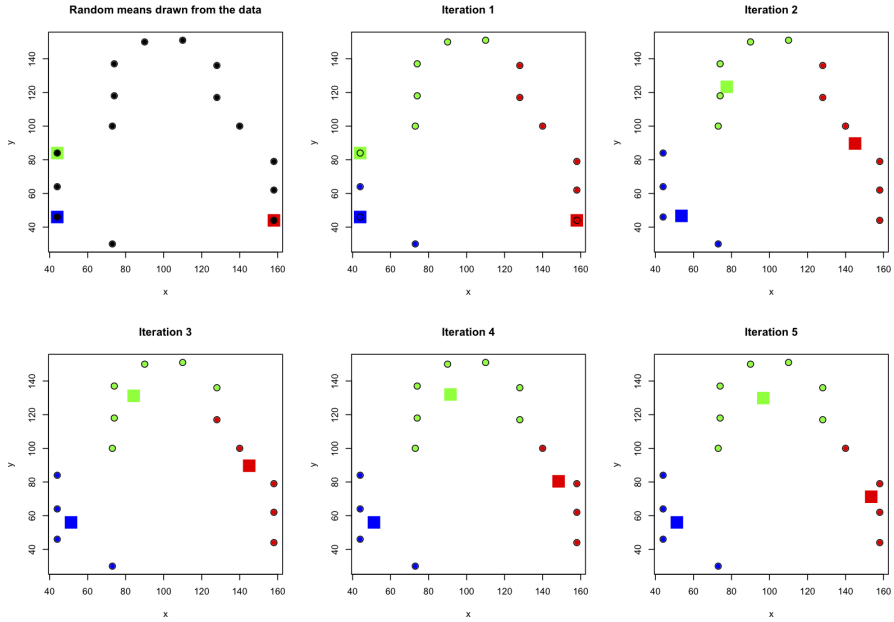

FIGURE 2.1 – Wikipedia k-means example

la figure 2.1 represente l'exemple de Wikipedia des k-means avec des valeurs numériques et des centroïdes choisis au hasard. Les centroïdes sont représentés par des carrés de couleur, et les points de données par des points, avec des couleurs qui représentes à quel cluster ils appartiennent. La première itération sert à affecter chaque point au cluster qui est le plus proche, et puis dans l'itération 2 en recalcule les nouveaux centroides on réaffectant les points aux nouveaux clusters et ainsi de suite pour chaque itération Dans les deux derniers panneaux, le changement des centroides ne modifie pas les clusters donc l'algorithme a convergé.

2.3.3 la classification supervisée

Consiste à examiner les caractéristiques d'un objet et lui attribuer une classe, et une opération qui permet de classer chaque individu dans sa classe, en fonction des caractéristiques de l'individu. La classe a des valeurs discrètes.[23],[2]

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 ces objets. Les méthodes s'étendent souvent à des variables Y quantitatives.

Nous disposons au départ d'un échantillon dit d'apprentissage dont le classement est connu. Cet échantillon est utilisé pour l'apprentissage des

règles de classement.la figure 2.2 représente un exemple de classification supervisée.

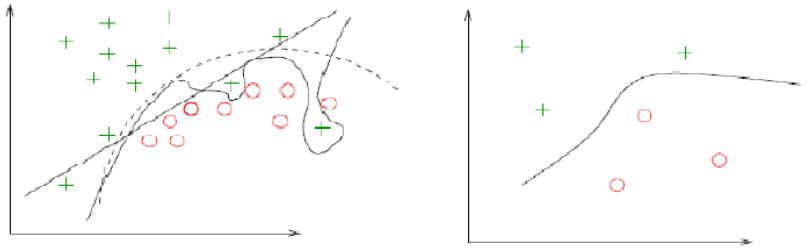

FIGURE 2.2 – classification supervisée

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). Nous utilisons souvent un deuxiéme échantillon indépendant, dit de validation ou de test.

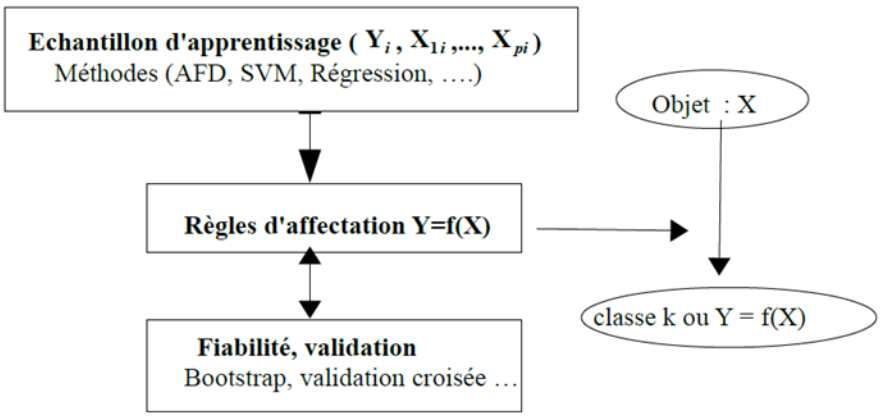

FIGURE 2.3 – démarche de classification supervisée

Parmi l'ensemble des méthodes de classification supervisée, nous allons parler sur Les machines à vecteurs de support ainsi que La Méthode des K-NN .

Les machines à vecteurs de support : les (SVM) sont une classe d'algorithmes d'apprentissage initialement définis pour la discrimination c'est-a-dire la prévision d'une variable qualitative initialement binaire. Ils ont été

ensuite généralisés à la prévision d'une variable quantitative. Dans le cas de la discrimination d'une variable dichotomique, ils sont basés sur la recherche de l'hyperplan de marge optimale qui, lorsque c'est possible, classe ou sépare correctement les données tout en étant le plus éloigné possible de toutes les observations. Le principe est donc de trouver un classifieur, ou une fonction de discrimination, dont la capacité de g''enéralisation (qualité de prévision) est la plus grande possible.[20] Le but de la classification par SVM est de trouver le séparateur dont la marge est la plus large possible, et cela sous les contraintes de séparation par les deux hyperplans.

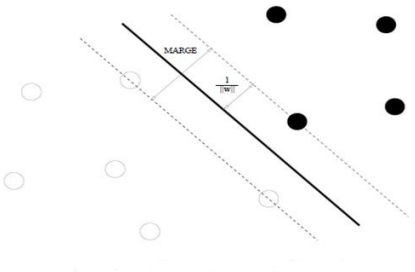

FIGURE 2.4 – marge d'un classifieur linéaire

La Méthode des K-NN La méthode des K plus proches voisins est une méthode d'apprentissage supervisée simple est efficace, en se basant sur la similarité entre les objets d'apprentissage (la base d'apprentissage), cette méthode est très adaptée aux problèmes multicritères[10] , le principe de cet algorithmes est le suivant :

- Calculer la distance entre l'objet à classer et l'ensemble des objets d'apprentissage ;
- Choisir les K plus proches objets ;
- Chaque objet effectue un « vote » pour la classe qu'il appartient, en incrémentant la variable associée ;
- Choisir la classe qui a eu le plus de vote.

Si S est le nombre de composantes représentant chaque individu ($S = n(resp.nXm)$ si l'individu est représenté par un vecteur (resp. une matrice)),T le nombre d'individus, et K le voisinage, La complexité du calcul des distances= $O(K \times S \times T)$.

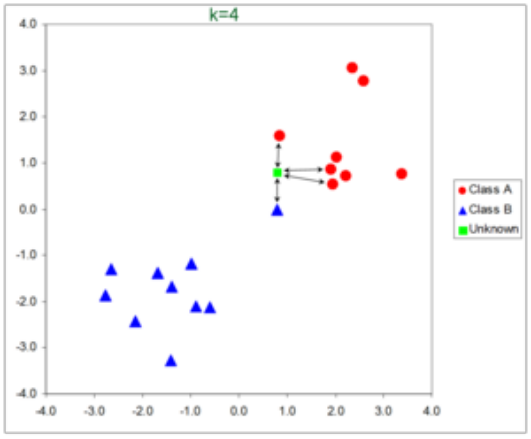

FIGURE 2.5 – knn example

Le calcul de similarité entre objets n'est pas toujours évident, on rencontre des problèmes liées aux structures des objets (linéaires, non linéaires, qualitatives, quantitatives, structurés, semi structurés, non structurés) à leurs tailles, à leurs descriptions ...
d'apres ce qu'on à défini precédement trois problèmes sont liées à la similarité qui sont les suivants :

2.4 problème

2.4.1 non linéarité

le but et de séparer chaque ensemble d'objets qui représente la même classe,Mais le problème est la distinction entre ces individus d'une manière linéaire,afin de pouvoir les séparer linéairement, car les problèmes naturels sont souvent non linéaires. On a donc une transformation d'un problème de séparation non linéaire dans l'espace de représentation en un problème de séparation linéaire dans un espace de re-description de plus grande dimension. Cette transformation non linéaire est réalisée via une fonction noyau.(voir figure 2.6

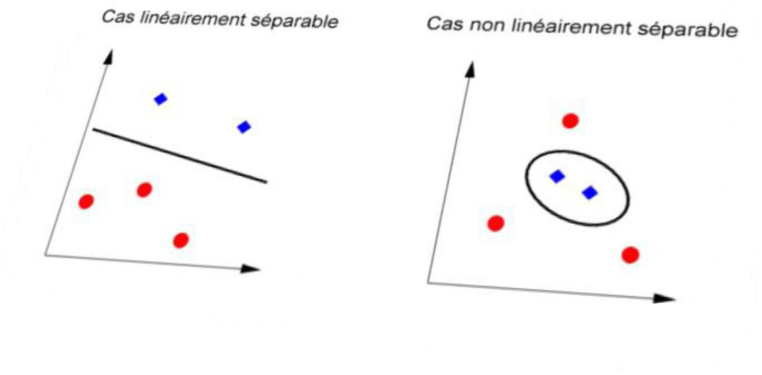

FIGURE 2.6 – linéarité et non linéarité

Nous constatons dans la figure 2.7 que la distinction entre le nuage bleu et rouge nécessite la définition d'une application (trait noir) qui n'est pas une fonction linéaire

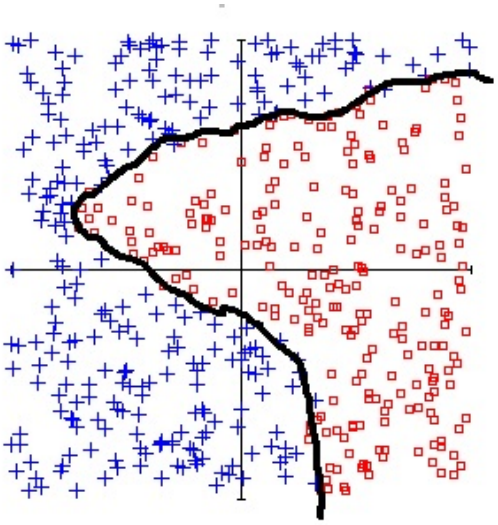

FIGURE 2.7 – objets non linéairement séparables

Une des solutions est d'utiliser les fonctions noyau qui sont utilisées pour permettre de distinguer linéairement deux ou plusieurs classes d'un nuage d'individus

la technologie des méthodes à noyau parait très utile dans les problèmes où les données sont hétérogènes et les données sont massives. En apprentissage

automatique, l'astuce du noyau (kernel trick) est une méthode qui consiste à utiliser un classifeur linéaire pour résoudre un problème non-linéaire, en transformant l'espace de représentation des données d'entrées en un espace de plus grande dimension,où le classifeur linéaire est alors utilisé. La discrimination linéaire dans l'espace de grande dimension (appelé aussi espace de redescription) est équivalente à une discrimination non-linéaire dans l'espace d'origine. Le kernel trick a été publié en 1964 par Aizerman et al[8] .
L'astuce du noyau s'utilise dans un algorithme qui ne dépend que du produit scalaire entre 2 vecteurs d'entrée x et y. Apres passage à un espace de redescription par une transformation,l'algorithme n'est plus dépendant que du produit scalaire :

$$< \theta(X), \theta(y) >$$

Le problème de ce produit scalaire est qu'il est effectué dans un espace de grande dimension, ce qui conduit a des calculs impraticables. L'idée est donc de remplacer ce calcul par une fonction noyau K telle que :

$$K(X, Y) =< \theta(X), \theta(y) >$$

2.4.2 taille variable

nous utilisons le même exemple pour présenter deux objets de structure et taille variable.
XML est l'exemple par excellence pour présenter des données semi structurées de taille variable, nous constatons que la figure 2.8 présente le document xml **DOC 1 :**

```
<repertoire>
    <!-- John DOE -->
    <personne sexe="masculin">
        <nom>DOE</nom>
        <prenom>John</prenom>
        <telephones>
          <telephone type="fixe">01 02 03 04 05</telephone>
          <telephone type="portable">06 07 08 09 10</telephone>
        </telephones>
    </personne>
</repertoire>
```

FIGURE 2.8 – structure de DOC1

la figure2.9 represente **l'Arbre de DOC1 :**

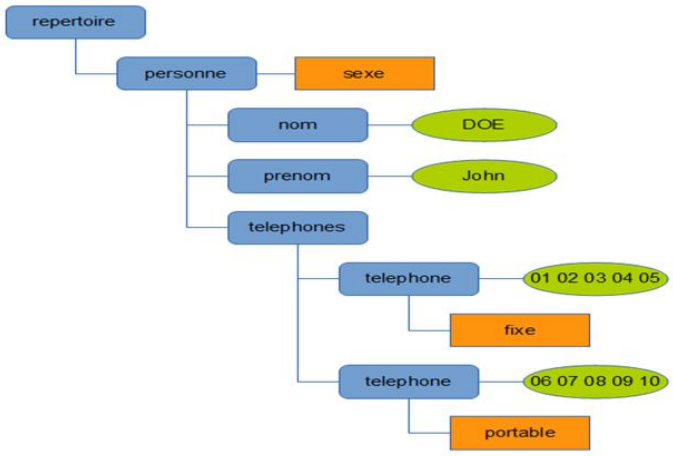

FIGURE 2.9 – l'arbre xml du DOC1

DOC1 est de Profondeur= 4

la figure 2.10 représente le document DOC2

```
<repertoire>
    <!-- John DOE -->
    <personne sexe="masculin">
        <nom>DOE</nom>
        <prenom>John</prenom>
        <telephones>01 02 03 04 05</telephones>
    </personne>
</repertoire>
```

FIGURE 2.10 – structure de DOC2

et la figure 2.11 représente son arbre.

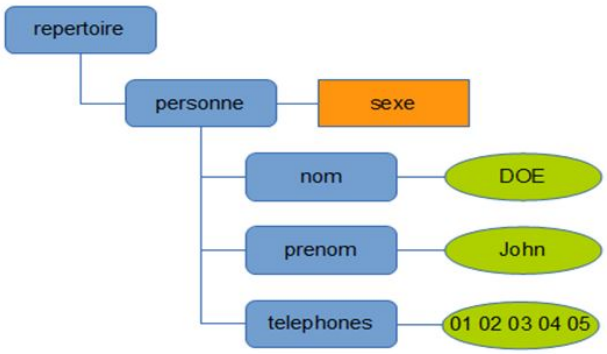

FIGURE 2.11 – l'arbre xml du DOC2

DOC2 est de Profondeur= 3

*DOC*1 et *DOC*2 sont similaires mais ne sont pas de même taille, comment calculer la distance entre ces documents ?
La méthode des n-grams, La régression et l'ACP sont des techniques utilisées pour résoudre ce genre de problème de calcul de distance, les deux derniéres techniques seront détaillées dans les prochains chapitres.

2.4.3 codage d'objet et quantité

si on a des objets présentés d'une manière naturelle(texte, fichiers bruts,...),des qualités (base de données,entrepôts de données,...)

Il faut tout d'abord trouver un bon codage :
- quantité
- taille fixe
- non ambiguë
- explicite
- ...

Le problème c'est que cette application n'est pas toujours envisageable. Pourquoi ne pas calculer dans l'espace de départ ?
Plusieurs techniques sont utilisées pour donner solution à ce problème comme :
- Les n grams pour les données textuelles, la bioinformatique et les bases de données.
- noyaux rationnels et transducteurs pour les langages de type 3

2.5 Conclusion

dans ce chapitre on a vu des généralités dur le calcul de similarité el le datamining.
et dans le suivant chapitre nous allons étudier quelque techniques statistiquesS pour remplacer les objets par des connaissances plus significatives " comportement" .

Chapitre 3

ACP et régression pour la distance

3.1 Introduction

Dans le chapitre précèdent, nous avons invoqué et étudié la notion de distance entre objets ainsi que les problèmes rencontrés.
Nous présentons dans ce chapitre deux méthodes génériques qui sont l'ACP, et la REGRESSION.et nous décrivons comment il est possible de les utiliser pour représenter les objets, et ainsi normaliser et résoudre les problèmes rencontrés.
En effet il existe beaucoup de techniques de normalisation, et de compression de données, nous intéressons dans ce mémoire de ces deux techniques dans le souhait d'étudier d'autres techniques dans le futur.

3.2 L'Analyse en Composante Principal (ACP)

Conçue par Karl Pearson en 1901, intégrée à la statistique mathématique par Harold Hotelling en 1933, l'analyse en composantes principales **(ACP)** n'est vraiment utilisée que depuis la large diffusion des moyens de calcul informatique.

La technique d'analyse en composantes principales peut être présentée de divers points de vue. Pour le statisticien classique, il s'agit de la recherche des axes principaux de l'ellipsoïde d'une distribution normale multidimensionnelle, ces axes étant estimés à partir d'un échantillon. C'est la présentation initiale de Hotelling (l933), puis celle des manuels classiques d'analyse multivariée.

Pour le factorialiste classique, il s'agit d'un cas particulier de la méthode d'analyse factorielle des psychométriciens (cas de variances spécifiques nulles ou égales).

Enfin, du point de vue plus récent des analystes de données, il s'agit d'une technique de représentation des données, ayant un caractère optimal selon certains critéres algébriques et géométriques, et que l'on utilise en général sans référence à des hypothèses de nature statistique ni à un modèle particulier.[24]

3.2.1 Définition informelle

L'Analyse en Composantes Principales (ACP) est une méthode d'analyse de données. Elle cherche à synthétiser l'information contenue dans un tableau croisant des individus et des variables quantitatives. Produire un résumé d'information au sens de l'ACP c'est établir une similarité entre les individus, chercher des groupes d'individus homogènes, mettre en évidence une typologie d'individus. Quant aux variables c'est mettre en évidence des bilans de liaisons entre elles, moyennant des variables synthétiques et mettre en évidence une typologie de variables. L'ACP cherche d'une façon générale à établir des liaisons entre ces deux typologies.[3]

En somme, les objectifs poursuivis par une ACP sont :

- La représentation graphique "optimale" des individus en minimisant les déformations du nuage des points, dans un sous-espace de dimension réduit.

– la représentation graphique des variables dans un sous-espace en explicitant au mieux les liaisons entre ces variables.

3.2.2 Domaines d'application

L'analyse en composantes principales est utilisée dans des diverses domaines parmi eux **la compression des données** et **les statistiques** qui comportent :
- **L'Analyse des données** qui est un ensemble de techniques descriptives, dont l'outil mathématique majeur est l'algèbre matricielle, et qui s'exprime sans supposer a priori un modèle probabiliste.
Elle comprend l'**ACP** qui est employée pour des données quantitatives.

- **la Visualisation des données** " représentation graphique de données statistiques "qui est un résumé visuel des données statistiques chiffrées. Elle permet en un seul coup d'œil d'en saisir la tendance générale.

- **ACP** est utilisé pour résoudre les problèmes de visualisation hors 3D.

l'ACP peut être utilisé comme une technique de compression irréversible dans le domaine de compression de données des images pour réduire la redondance des données d'une image afin de pouvoir l'emmagasiner sans occuper beaucoup d'espace ou la transmettre rapidement.

3.2.3 Définition formelle de l'ACP

L'étude d'une population statistique de taille n passe le plus souvent par le recueil d'un nombre élevé p de données quantitatives par élément observé. L'analyse de ces données doit tenir compte de leur caractère multidimensionnel et révéler les liaisons existantes entre leurs composantes [11].

L'analyse en composantes principales (ACP), est une méthode très puissante pour explorer la structure de telles données. Chaque donnée étant représentée dans un espace à p dimensions. L'ensemble des données forme un nuage de n points dans Rp. Le principe de l'ACP est d'obtenir une représentation approchée du nuage dans un sous-espace de dimension faible k par projection sur des axes bien choisis. Une métrique dans Rp étant choisie (en général normalisée par l'utilisation de variables centrées réduites). Les k axes principaux sont ceux qui maximisent l'inertie du nuage projeté, c'est-à-dire la moyenne pondérée des carrés des distances des points projetés à leur centre

de gravité. Les composantes principales sont les n vecteurs ayant pour coordonnées celles des projections orthogonales des n éléments du nuage sur les k axes principaux [25] L'ACP construit ainsi de nouvelles variables artificielles, et des représentations graphiques permettant de visualiser les relations entre variables, ainsi que l'existence éventuelle de groupes d'éléments et de groupes de variables.

Soit Y la matrice des individus

$$Y = \begin{bmatrix} y_1^1 & ... & y_1^2 & ... & y_1^p \\ y_2^1 & ... & y_2^2 & ... & y_2^p \\ ... & & & & \\ y_n^1 & ... & y_n^2 & ... & y_n^p \end{bmatrix}$$

La matrice centrée $\bar{Y}$ est calculée comme suit :

$$\bar{Y} = \begin{bmatrix} y_1^1 - \bar{Y}_1 & ... & y_1^2 - \bar{Y}_2 & ... & y_1^p - \bar{Y}_p \\ y_2^1 - \bar{Y}_1 & ... & y_2^2 - \bar{Y}_2 & ... & y_2^p - \bar{Y}_p \\ ... & & & & \\ y_n^1 - \bar{Y}_1 & ... & y_n^2 - \bar{Y}_2 & ... & y_n^p - \bar{Y}_p \end{bmatrix}$$

avec $\bar{Y}_i$ est la moyenne de la colonne i.

la matrice reduite $\tilde{Y}$ est calculée comme suit :

$$\tilde{Y} = \begin{bmatrix} (y_1^1 - \bar{Y}_1)/\sigma(Y_1) & ... & (y_1^2 - \bar{Y}_2)/\sigma(Y_2) & ... & (y_1^p - \bar{Y}_p)/\sigma(Y_p) \\ (y_2^1 - \bar{Y}_1)/\sigma(Y_1) & ... & (y_2^2 - \bar{Y}_2)/\sigma(Y_2) & ... & (y_2^p - \bar{Y}_p)/\sigma(Y_p) \\ ... & & & & \\ (y_n^1 - \bar{Y}_1)/\sigma(Y_1) & ... & (y_n^2 - \bar{Y}_2)/\sigma(Y_2) & ... & (y_n^p - \bar{Y}_p)/\sigma(Y_p) \end{bmatrix}$$

avec σY_i représente l'écart type de la colonne i.

Une fois la matrice Y transformée en $\bar{Y}$ ou $\tilde{Y}$,il suffit de la multiplier par sa transposée pour obtenir :
$\bar{M}$ la matrice de variances-covariances si Y est centrée,
$\tilde{M}$ la matrice de corrélation si Y est réduite.

Finalement, nous cherchons le vecteur u tel que la projection du nuage sur u ait une variance maximale. Cette projection s'écrit :

$$\pi_u(\bar{M}) = \bar{M}.u(respectivement\tilde{M})$$

Le vecteur u représente donc le vecteur propre associé à la valeur propre $\lambda 1$.

La diagonalisation de la matrice de corrélation (ou de covariance si on se place dans un modèle non réduit), nous a permis d'établir que le vecteur qui explique le plus l'inertie du nuage, est le premier vecteur propre. De même le deuxième vecteur qui explique la plus grande part de l'inertie restante est le deuxième vecteur propre, etc.

3.2.4 Exemlpe ACP

Considérons pour l'exemple une étude d'un botaniste qui a mesuré les dimensions de 15 fleurs d'iris. Les trois variables equation mesurées sont :

Fleur n°	x_1	x_2	x_3
1	5.1	3.5	1.4
2	4.9	3.0	1.4
3	4.7	3.2	1.3
4	4.6	3.1	1.5
5	5.0	3.6	1.4
6	7.0	3.2	4.7
7	6.4	3.2	4.5
8	6.9	3.1	4.9
9	5.5	2.3	4.0
10	6.5	2.8	4.6
11	6.3	3.3	6.0
12	5.8	2.7	5.1
13	7.1	3.0	5.9
14	6.3	2.9	5.6
15	6.5	3.0	5.8

FIGURE 3.1 – Exemple pratique de données tabulaires ACP

Pour nous un tel tableau de données sera tout simplement une matrice réelle a n lignes (les individus) et a p colonnes (les variables)
calculons la matrice centrée,on obtient alors4.23 :

$$X_c = \begin{bmatrix} -0.8067 & 0.4400 & -2.4733 \\ -1.0067 & -0.0600 & -2.4733 \\ -1.2067 & 0.1400 & -2.5733 \\ -1.3067 & 0.0400 & -2.3733 \\ -0.9067 & 0.1400 & -2.4733 \\ 1.0933 & 0.5400 & 0.8267 \\ 0.4933 & 0.1400 & 0.6267 \\ 0.9933 & 0.0400 & 1.0267 \\ -0.4067 & -0.7600 & 0.1267 \\ 0.5933 & -0.2600 & 0.7267 \\ 0.3933 & 0.2400 & 2.1267 \\ -0.1067 & -0.3600 & 1.2267 \\ 1.1933 & -0.0600 & 2.0267 \\ 0.3933 & -0.1600 & 1.7267 \\ 0.5933 & -0.0600 & 1.9267 \end{bmatrix}$$

FIGURE 3.2 – matrice centrée

Pour donner une importance identique a chaque variable afin que le type d'unités des mesures n'influence pas l'analyse, nous travaillerons avec les données centrées réduites. Pour cela,on aura :

$$Z = \begin{bmatrix} -0.2383 & 0.3572 & -0.3375 \\ -0.2974 & -0.0487 & -0.3375 \\ -0.3565 & 0.1136 & -0.3512 \\ -0.3861 & 0.0324 & -0.3239 \\ -0.2679 & 0.4384 & -0.3375 \\ 0.3230 & 0.1136 & 0.1128 \\ 0.1457 & 0.1136 & 0.0855 \\ 0.2935 & 0.0324 & 0.1401 \\ -0.1201 & -0.6170 & 0.0172 \\ 0.1753 & -0.2110 & 0.0991 \\ 0.1162 & 0.1948 & 0.2902 \\ -0.0315 & -0.2922 & 0.1674 \\ 0.3526 & -0.0487 & 0.2765 \\ 0.1162 & -0.1298 & 0.2356 \\ 0.1753 & -0.0487 & 0.2629 \end{bmatrix}$$

FIGURE 3.3 – matrice centree reduite

le calcul de La matrice des variances-covariances donne pour notre exemple, la matrice carrée suivante :

$$R = \begin{bmatrix} 1 & -0.1609 & 0.8854 \\ -0.1609 & 1 & -0.3818 \\ 0.8854 & -0.3818 & 1 \end{bmatrix}$$

FIGURE 3.4 – La matrice des variances-covariances

les trois valeurs propres de la matrice de corrélation R sont :

$$\lambda_1 = 2.0303 \quad \lambda_2 = 0.8846 \quad \lambda_3 = 0.0853$$

$$\Lambda = \begin{bmatrix} 2.0303 & 0 & 0 \\ 0 & 0.8846 & 0 \\ 0 & 0 & 0.0853 \end{bmatrix}$$

FIGURE 3.5 – la matrice de correlation

$$P_{\lambda 1} = \frac{2.0303}{\sum_i \lambda_i} = 66.67\% \quad P_{\lambda 2} = \frac{0.8846}{\sum_i \lambda_i} = 29.48\% \quad P_{\lambda 3} = \frac{0.0853}{\sum_i \lambda_i} = 2.84\%$$

FIGURE 3.6 – valeurs propres

Donc la premiere composante explique 66.67% de l'effet. Les deux premieres composantes en expliquent 96.15%, etc. C'est ainsi par exemple en finance que l'on va prendre parmi une dizaine ou plus de composantes, seulement celles qui mènent a "expliquer" le 95%.

En ayant les trois valeurs propres, pour déterminer les trois vecteurs propres $(\overrightarrow{u_1}, \overrightarrow{u_2}, \overrightarrow{u_3}$ qui forment la base principale, il nous faut donc resoudre le systeme de trois equations a trois inconnues suivant pour chaque valeur propre :

$$\underbrace{\begin{bmatrix} 1-\lambda_i & -0.1609 & 0.8854 \\ -0.1609 & 1-\lambda_i & -0.3818 \\ 0.8854 & -0.3818 & 1-\lambda_i \end{bmatrix}}_{R} \underbrace{\begin{bmatrix} u_{i,1} \\ u_{i,2} \\ u_{i,3} \end{bmatrix}}_{\vec{u_i}} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

FIGURE 3.7 – systeme d'equations

.

.

Ce qui donne donc la matrice des vecteurs propres :

$$S = (\vec{u}_1, \vec{u}_2, \vec{u}_3) = \begin{bmatrix} 0.6410 & 0.3803 & -0.6666 \\ -0.3527 & 0.9174 & 0.1841 \\ 0.6816 & 0.1171 & 0.7222 \end{bmatrix}$$

FIGURE 3.8 – la matrice des vecteurs propres

Nous avons alors comme coordonnees des points M_i dans la base $(\overrightarrow{u_1}, \overrightarrow{u_2}, \overrightarrow{u_3}$ en utilisant :

$$\psi = Z.S$$

la matrice suivante :

$$\psi = \begin{bmatrix} -0.5268 & 0.2045 & -0.0198 \\ -0.4178 & -0.2043 & -0.0564 \\ -0.5259 & -0.0750 & 0.0052 \\ -0.4966 & -0.1604 & 0.0305 \\ -0.5760 & 0.2699 & 0.0161 \\ 0.2525 & 0.2488 & -0.1169 \\ 0.1156 & 0.1757 & -0.0150 \\ 0.2818 & 0.1634 & -0.0916 \\ 0.1578 & -0.6311 & -0.0218 \\ 0.2634 & -0.1194 & -0.0871 \\ 0.2108 & 0.2660 & 0.1739 \\ 0.2039 & -0.2697 & 0.0912 \\ 0.4469 & 0.1261 & -0.0459 \\ 0.2908 & -0.0490 & 0.0712 \\ 0.3196 & 0.0546 & 0.0663 \end{bmatrix}$$

FIGURE 3.9 – matrice des coordonnees des projections

Les coordonnees des projections du nuage de points dans le meilleur plan defini par les vecteurs$(\overrightarrow{u_1}, \overrightarrow{u_2})$ sont donc les deux premieres colonnes de la

matrice precedente (correspondant donc a la longueur du sepale et la largeur du sepale).

Effectivement, nous voyons immédiatement que ce sont ces deux colonnes qui maximiseront la somme des normes dans le plan donné :

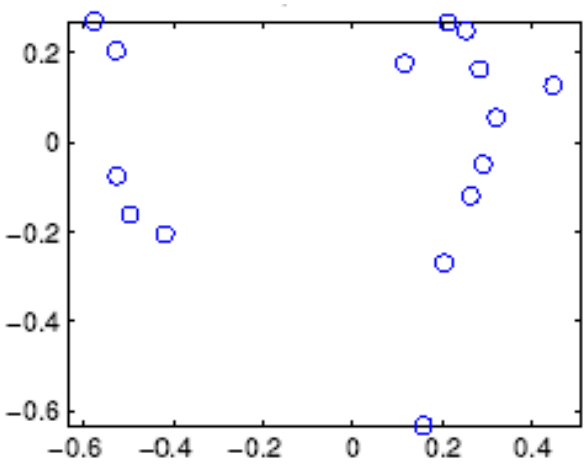

FIGURE 3.10 – Plan factoriel

3.3 la régression :méthode des moindres carrées

La régression ou **la méthode des moindres carrés**, indépendamment élaborée par Legendre et Gauss au début du XIX^e siècle, permet de comparer des données expérimentales, généralement entachées d'erreurs de mesure, à un modèle mathématique censé décrire ces données.

3.3.1 Définition informelle

La régression est un ensemble de méthodes statistiques très utilisées pour analyser la relation d'une variable par rapport à une ou plusieurs autres.
Pendant longtemps, la régression d'une variable aléatoire Y sur le vecteur de variables aléatoires X désignait la moyenne conditionnelle de Y sachant X. Aujourd'hui, le terme de régression désigne tout élément de la distribution conditionnelle de Y sachant X considérée comme une fonction de X .
On peut par exemple s'intéresser à la moyenne conditionnelle, à la médiane conditionnelle, au mode conditionnel, à la variance conditionnelle...[15]

3.3.2 Domaine d'application

- **Analyse discriminante :**Elle sert essentiellement à décrire ce qui distingue les moyennes par groupe de plusieurs variables mesurées sur plusieurs individus de plusieurs groupes.
- **Approximation statistique :**utiliser pour trouver une courbe approximative à un ensemble de points de mesures empiriques.
- **Classification binaire :**la courbe de regression sert de délimiteur binaire entre deux familles de points

3.3.3 Définition formelle

soit X un ensemble d'étiquettes et Y un ensemble d'observations et $f(\psi, x)$ la fonction de regression, la methode des moindres carrées consiste a minimiser la valeur suivante :

$$ArgMin \underset{\sigma \in \psi}{[} y_i - [f(\psi, x_i)^2]]$$

3.3.4 La régression linéaire

Une situation courante en sciences biologiques est d'avoir à sa disposition deux ensembles de données de taille $n, y_1, y2, ..., y_n$ et $x_1, x_2, ..., x_n$, obtenus expérimentalement ou mesures sur une population. Le problème de la régression consiste à rechercher une relation pouvant éventuellement exister entre les x et les y, par exemple de la forme $y = f(x)$. Lorsque la relation recherchée est affinée, c'est-a-dire de la forme $y = ax + b$, on parle de régression linéaire. Mais même si une telle relation est effectivement présente, les données mesurées ne vérifient pas en général cette relation exactement. Pour tenir compte dans le modèle mathématique des erreurs observées, on considère les données $y_1, y_2, ..., y_n$ comme autant de réalisations d'une variable aléatoire Y

et parfois aussi les données $x_1, x_2, ..., x_n$ comme autant de réalisations d'une variable aléatoire X. On dit que la variable Y est la variable dépendante ou variable expliquée et que la variable X est la variable explicative [16]

La droite des moindres carrées : Les données $(x_i; y_i)\ ; i = 1, ..., n$ peuvent être représentées par un nuage de n points dans le plan. Le centre de gravité de ce nuage $(\bar{x}; \bar{y})$ peut se calculer comme suit :

$$(\bar{x}, \bar{y}) = \frac{1}{n}(\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} y_i)$$

Rechercher une relation affine entre les variables X et Y revient à déterminer une droite qui s'ajuste le mieux possible à ce nuage de points. Parmi toutes les droites possibles, on retient celle qui jouit d'une propriété remarquable : c'est celle qui rend minimale la somme des carrées des écarts des valeurs observées yi à la droite

$$by_i = ax_i + b$$

. Si ϵi représente cet écart, appelé aussi r''esidu, le principe des moindres carrés ordinaire (MCO) consiste à choisir les valeurs de a et de b qui minimisent

$$E = \sum_{i=0}^{n} \epsilon_i^2 = \sum_{i=0}^{n} (y_i - (ax_i + b))^2$$

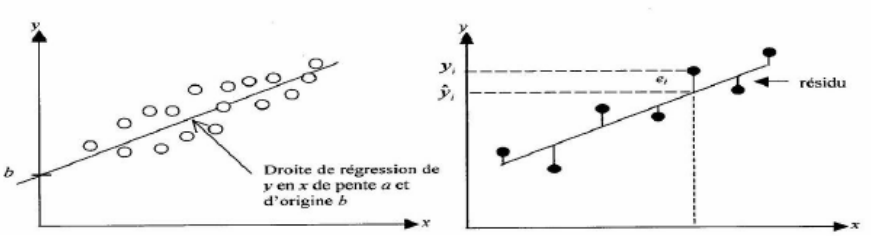

FIGURE 3.11 – droite des moindres carrées

Il est facile de montrer que les valeurs notées $\hat{a}$ et $\hat{b}$ sont respectivement égales à

$$\hat{a} = \frac{\sum_{i=1}^{n}(x-\bar{x})(x-\bar{x})}{\sum_{i=1}^{n}(y-\bar{y})^2}$$

$$\hat{b} = \bar{y} - \hat{a}(x)$$

Et. On exprime souvent $\hat{a}$ au moyen de la variance et de la covariance des variables aléatoires X et Y par :

en effet on a :

Pour mesurer la qualité de l'approximation d'un nuage $(x_i; y_i)$ par sa droite des moindres carrées, on calcule le coefficient de corrélation linéaire défini par :

$$r_{xy} = \frac{cov_x y}{s_x s_y}$$

Notons que,la droite de régression passe par le barycentre du nuage de points,qui a pour coordonnées $G(\bar{x}, \bar{y})$.

exemlpe de regression linéaire

Considérons la figure 4.19. Nous avons quatre points expérimentaux, de coordonnées :

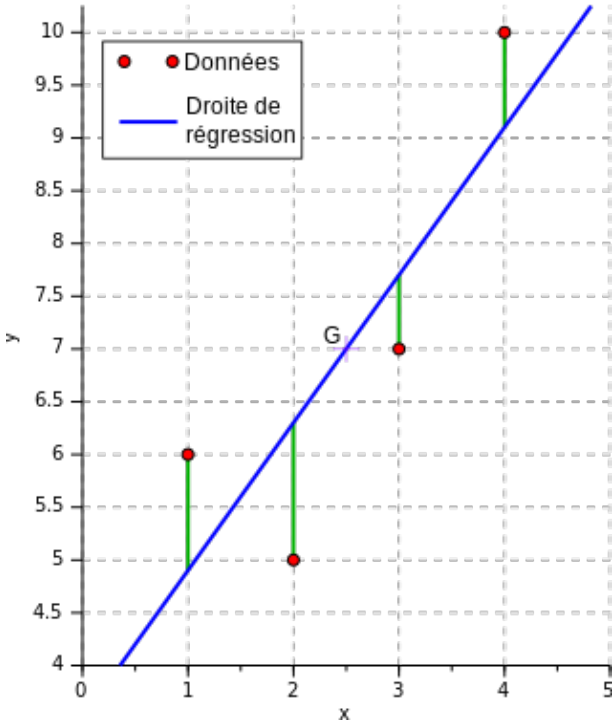

FIGURE 3.12 – exemple de regression linéaire

- $p_1(1;6)$
- $p_2(2;5)$
- $p_3(3:7)$
- $p_4(4:10)$

Nous calculons :

$$\bar{x} = \frac{1+2+3+4}{4} = 2,5$$

$$\bar{y} = \frac{6+5+7+10}{4} = 7$$

Nous savons donc que la droite de régression passe par le point $G(2,5;7)$.
Et donc

$$cov(x) = \frac{(1-2,5)(6-7)+(2-2,5)(5-7)+(3-2,5)(7-7)+(4-2,5)(10-7)}{4} = \frac{7}{4} = 1,7$$

$$S_x^2 = var(x) = \frac{(1-2,5)^2+(2-2,5)^2+(3-2,5)^2+(4-2,5)^2}{4} = \frac{5}{4} = 1,25$$

$$\hat{a} = \frac{1,75}{1,25} = 1,4$$

$$\hat{b} = 7 - 1,4 \times 2,5 = 3,5$$

L'équation de la droite de regression est donc $y = 1,4x + 3,5$ Nous pouvons comparer les valeurs experimentales de y avec les valeurs calculées (sur la droite de regression)

i	x	y_{exp}	y_{cal}	u
1	1	6	4,9	1,1
2	2	5	6,3	-1,3
3	3	7	7,7	-0,7
4	4	10	9,1	0,9

FIGURE 3.13 – valeurs calculées

Par ailleurs, si l'on suppose (hypothèses du Theoreme de Gauss-Markov) que les variables aleatoires U_i :

- sont centrées (hypothèse d'exogénéité) : $E(U_i) = 0$;
- ont la même variance (hypothèse d'homoscédasticité) ;
- sont indépendantes (hypothèse de non-corrélation) $E(U_i U_j)_i \neq j = 0$;

alors on a :

$$\sum_{i=1}^{n} u_i^2 = n(var(U) + \bar{u}^2)$$

où :

- $\bar{u}$ est la moyenne empirique des $Ui, \bar{u} = \frac{1}{n}\sum u_i$
- var est la variance empirique,var(U) $= \frac{1}{n}\sum(u_i - \bar{u})^2$.

On a :

- $var(X) \neq 0$;
- $var(Y) \neq 0$;
- $var(X)var(Y) \geq cov^2(X, Y)$ le produit des variances est supérieur ou égal au carré de la Covariance.

donc si l'on pose

$$r_{xy} = \frac{\mathrm{cov(X, Y)}}{\sqrt{\mathrm{var(X)\,var(Y)}}} = \frac{\mathrm{cov(X, Y)}}{S_x S_y}$$

on a

$$-1 \leq r_{xy} \leq 1.$$

$$\text{var(Y)} = \frac{(6-7)^2 + (5-7)^2 + (7-7)^2 + (10-7)^2}{4} = \frac{14}{4} = 3,5$$

$$r_{xy} = \frac{1,75}{\sqrt{1,25 \times 3,5}} \simeq 0,8367$$

3.3.5 La régression non linéaire :

Il existe plusieurs modèles d'ajustement non linéaires parmi lesquels on cite un modèle qui est l'ajustement parabolique de la forme $y = a_0 + a_1 x + a_2 x^2$. On peut réaliser cet ajustement non linéaire par le principe de la méthode des moindres carrés. On cite un exemple d'ajustement parabolique.
Soit la série statistique donnée par le tableau ci- dessous.
série statistique donnée par le tableau ci- dessous.

X	1	3	4	5
Y	1	4	2	0

FIGURE 3.14 – tableau de serie statistique

Notre but est de faire un ajustement parabolique de la forme :
$y = a_0 + a_1 x + a_2 x^2$
sur cet exemple On obtient un système de trois équations à trois inconnus.

$$\sum y_i = n a_0 + a_1 \sum x_i + a_2 \sum x_i^2$$

$$\sum x_i y_i = a_0 \sum x_i + a_1 \sum x_i^2 + a_2 \sum x_i^3$$

$$\sum x_i^2 y_i = a_0 \sum x_i^2 + a_1 \sum x_i^3 + a_2 \sum x_i^4$$

Apres avoir effectué les calculs on obtient le systeme suivant :

$$7 = 4a_0 + 13a_1 + 51a_2$$

$$21 = 13a_0 + 51a_1 + 217a_2$$

$$69 = 51a_0 + 217a_1 + 963a_2$$

On peut résoudre ce système par plusieurs méthodes (méthodes des matrices ou pivot de GAUSS ou CRAMER).On résout ce système par la méthode de CRAMER méthode des déterminants .On trouve une solution unique qui est :

$$a_0 = -2.60; a_1 = 4.46 et a_2 = -0.80$$

D'où l'equation de l'ajustement parabolique est :$y = -0.80x^2 + 4.46x - 2.60$
Remarque : cette parabole ne passe pas par le point moyen $G(\bar{X}, \bar{Y})$ le centre de gravité des points.

3.4 Conclusion

Dans ce chapitre nous avons détaillés la technique de régression ainsi que l'ACP en donnant les formules de calculs illustrées par des exemples.et dans le suivant chapitre nous allons effectuer une comparaison entre ces deux derniéres.

Chapitre 4

étude comparative

4.1 Introduction

Nous avons vu quelques notions sur les distances. Nous nous contenons d'étudier deux techniques parmi elles : la régression et l'ACP.

Nous nous intéressons dans un premier temps d'effectuer une étude comparative entre les deux techniques, leurs implémentations..., dans un second temps nous étudions quelques issues d'implementation et nous testons quelques problèmes connus "arbre XML, cube de données"

4.2 Contexte

Le domaine d'application de l'ACP est très vaste et la mme chose pour la régression, pour qu'on peut comparer ces deux techniques on doit bien définir un contexte. le cadre de notre comparaison se portera sur :

- Le calcul de distance
- l'aspect rapproché des deux techniques. il y a des pertes d'informations.
- l'aspect multicritère des deux techniques : sur le mono critère, l'ACP ne fonctionne pas

4.3 Comparaison"ACP Vs Régression"

4.3.1 Quand utiliser l'ACP ?

L'analyse en composante principal est usuellement utilisée :

- comme outil de compression **exemple :** compression des images satellitaires
- dans l'analyse des données
- dans la projection/visualisation : projeter les données pour la visualisation hors 3D (trois dimensions)
- dans les calculs des distances
- l'ACP doit obligatoirement utilisée dans un domaine multicritères
- sur des données non creuses : peu d'absence d'information d'individus/variables

4.3.2 Quand utiliser la régression ?

- Comme l'ACP la régression est aussi utilisée dans la compression des données
- l'approximation
- La régression s'applique dans un espace défini, et peu importe creux ou non
- Elle a la souplesse d'étudier n'importe quel phénomène
- La régression peut s'appliquer au monocritère

4.3.3 Complexité

La complexité de l'ACP est quadratique $\theta(n^2)$ et nécessite des multiplications des matrices. En effet on constate que le calcul des matrices variance-covariance, corrélation nécessite une multiplication des matrices de départ, une opération qui coute $\theta(n^2)$ néanmoins, le calcul des vecteurs, valeurs

propres peut se faire en $\theta(n \log n)$;la complexité globale sera donc dans l'ordre de $\theta(n^2)$

La complexité de la régression peut varier selon la fonction linéaire utilisée, sa dépend juste de deux paramétres a et b, le calcul est fait en $\theta(n \log n)$ elle est généralement peu complexe, sauf dans les fonctions complexes (fonctions topologiques qui sont 2^n)

4.3.4 Données qualitatives

L'ACP ainsi que la régression nécessite une fonction d'évaluation (de coût / d'utilité)

4.3.5 Données semi et non structurées

l'ACP nécessite une application de l'ensemble de structure vers le power set des réels :$S \rightharpoonup R^n$

Prenant un exemple illustratif :Soit l'ensemble de départ S1 qui contient les deux documents XML ,xml1 et xml2

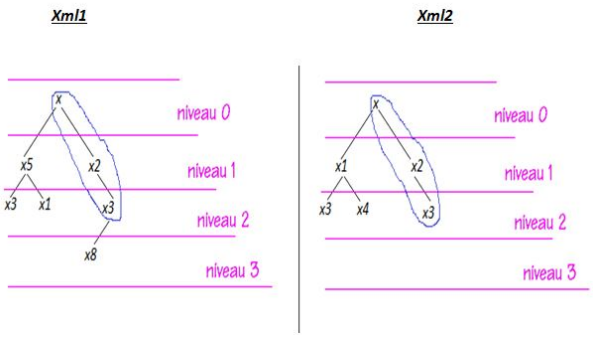

FIGURE 4.1 – arbres xml similaires

Premiérement on effectue un étiquetage et on s'assure qu'il soit unique, La solution est d'utiliser un étiquetage fermé pour les préfixes(voir figure 4.2 :

1. $\forall uv \in E, u \in E$ (fermé pour les contextes)
2. $\forall uv \in E, j \in N, ui \in E, 0 \leq i \leq j$ (ordre)

exemple d'étiquetage

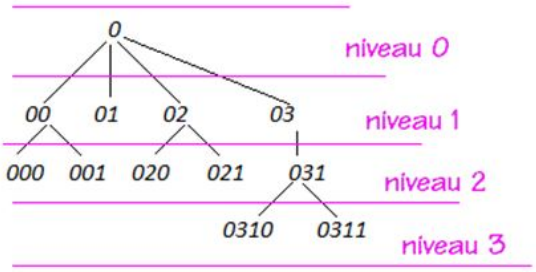

FIGURE 4.2 – exemple d'etiquetage

On applique l'étiquetage sur les deux documents XML, on obtient la figure 4.3 :

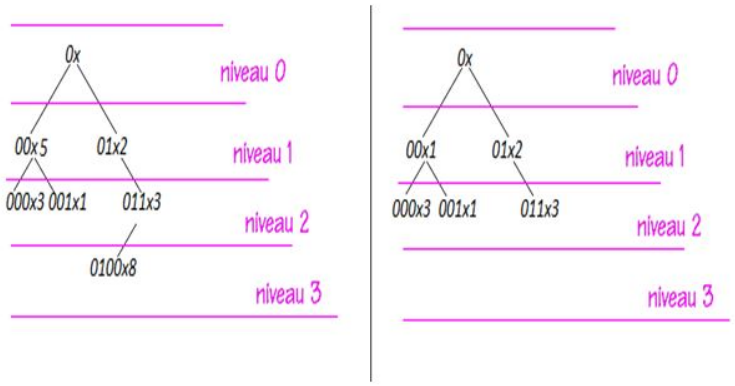

FIGURE 4.3 – etiquetage appliqué sur les deux arbres

On extrait les deux représentations uniques qui appartiennent à l'ensemble $S2$, on obtient 4.4

Xml1	0x	00x5	01x2	000x3	001x1	011x3	0110x8

Xml2	0x	00x1	01x2	000x3	001x4	011x3

FIGURE 4.4 – representation des deux doc

On remarque que les deux représentations ne sont rien qu'une application $E \rightharpoonup < XML >$ tel que $< XML >$ représente les noeuds d'un document XML

Si on définit un semi anneau transitif $(S, \bigoplus, \bigotimes)$ muni de deux lois de compositions $\bigoplus$ et $\bigotimes$, et on définit une relation d'ordre $\preceq$ qui peut comparer les éléments de S. l'application de moindre carrées peut s'étendre sur ce semi anneau, et la définition d'une application bijective sera plus évidente
Sur l'exemple suivant, la transformation $\phi(S1 \rightharpoonup S2)$ est une bijection.
Contrairement à l'ACP qui nécessite un calcul de vecteur, valeur propre (calcul linéaire)
On doit transformer S vers R^n
Sur l'exemple suivant, on propose l'application suivante
Nous avons trouvé une application $S1 \rightharpoonup S2$

$X1 \rightharpoonup 1$

$X2 \rightharpoonup 2$

$$\varphi \circ \varphi'$$

$X3 \rightharpoonup 3$

$$(S1 \rightharpoonup S2)(S2 \rightharpoonup R^* * R)$$

...

$S \xrightarrow{\varphi 1 \circ \varphi 2} R^* * R$

Pour éviter le problème de l'étoile * On calcule $\varphi'' : R^* * R \rightharpoonup R^m * R$
Tel que m est la taille max d'une représentation dans $S2$, on aura une représentation $R^m * R$ pour tout document XML, l'application de l'ACP devient donc évidente

Problème de bijection est-ce que $(\varphi \circ \varphi' \circ \varphi'')$?
Il est très difficile de prouver la bijection vu les problèmes :

- Transformation basée sur le
- Taille Gigant espace

L'ACP peut rencontrer des problèmes si :

- **L'application n'est pas bijective :** généralement, les données utilisées dans les datamining sont qualitatives, non uniforme, non réduite(échelles non normalisés)
  Le besoin de transformer ses donnéen quantité n'est pas évitable.de plus si la donnée est semi ou non structurée, une transformation de plus doit être appliquée pour présenter les données sous forme linéaire. Si S est la donnée φ est la première transformation qui permet de plonger S vers une autre representation lineaire S' φ transforme S' en quantités s'il n'y a pas de regles de transformation, et la definition de φ et φ' est laissé a l'utilisateur, ce qui pose un probleme de bijectivité , une application ainsi defini ne peut etre obligatoirement bijective.
- **Manque de preuve** : une consequence du premier point est la difficulté de prouver que $\phi \circ \phi'$ est bijective a savoir definissable, calculable, denombrable, decidable
- **Complexité** concernant la complexité, le calcul de matrices de variance-covariance (corrélation) nécessite des multiplications de matrice qui est une tache quadrazque.si n est la taille de la donnée, ACP tourne autour de()

Quant à La régression :

- La notion de $+, \times \rightarrow$ semi anneau
- Ordre $\rightarrow$ les ensembles
- Dérivé $\rightarrow$ presente dans les ensembles

On peut la généraliser aux espaces non classiques

la figure 4.5 représente le tableau comparatif "ACP Vs Regression" qui resume notre étude comparative

	ACP	Régression
Quand utiliser	- Multicritère - Compression - Calcul de distance - Non ou peu creuse	- Mono/multicritère - Compression - Calcul de distance - Peu importe
Complexité	- Quadratique $O(n^2)$ - Problème en temps réel - Beaucoup de calcul	- Selon la fonction, généralement quasi-linéaire $O(n\log(n))$ - Adapté au temps réel - Moins de calcul
Données qualitatives	- Fonction d'évaluation	- Fonction d'évaluation
Données semi et non structurées	- Filtre d'application	-peut s'adapter au multicritère

FIGURE 4.5 – tableau comparatif

4.4 Étude de cas

On va faire une étude de cas sur des exemples de données structurées pour des entrepots de données, et des documents XML
Le choix s'est fait aléatoirement sur ses deux cas. D'autre cas seront visés dans des travaux futurs

4.4.1 cas1 :Document XML

Définition

La norme XML (eXtensible Markup Language) décrit simplement comment construire un fichier texte permettant de stocker des informations en respectant une structure donnée. On parle alors de document XML[27]
Il existe deux méthodes essentielles appelées **SAX** et **DOM** pour lire un document XML dans un fichier

SAX (Simple Api for XML)

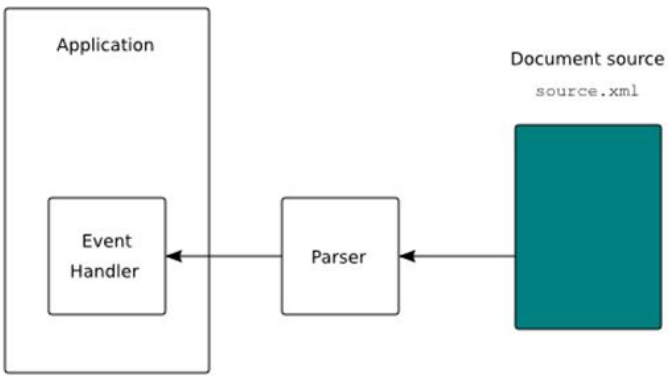

FIGURE 4.6 – structure SAX

SAX est une API permettant de lire un fichier XML sous forme de flux. Le principe de fonctionnent est le suivant. L'application crée un parseur et elle enregistre auprès de ce parseur son gestionnaire d'événements. Au cours de la lecture du fichier contenant le document XML (voir figure4.6, le gestionnaire reçoit les evenements générés par le parseur. Le document XML n'est pas chargé en mémoire.[13]

DOM (Document Object Model)

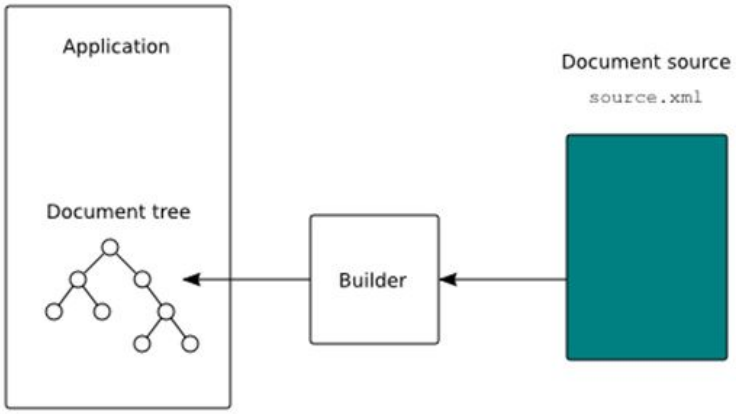

FIGURE 4.7 – structure DOM

DOM est une API permettant de charger un document XML sous forme d'un arbre qu'il est ensuite possible de manipuler. Le principe de fonctionnent est le suivant.(voir figure4.7 L'application crée un constructeur qui lit le document XML et construit une representation du document XML sous forme d'un arbre.[13]

Proposition

Nous proposons une réduction simple afin de transformer un document XML en une structure qu'on peut utiliser ACP ou bien la regression dessous. Nous reproduisons la notion d'étiquetage présenté dans 4.4.
Enfin nous aurons une représentation de la forme $< XMLnode > \times R^n$
Où $< XMLnode >$ est l'ensemble des noeuds du document et n la profondeur de l'arbre
Ensuite nous trions $< XMLnode >$ un tri de chaine de caractere, et on applique la transformation suivante Pour l'ACP, afin de garder une présentation multidimensionnelle : $n \in < XMLnode >$

$$\phi(x) = n, \forall y > x : \phi(y) > \phi(x)$$

donc en obtient la figure 4.8suivante :

Xml1	01	006	013	0004	0012	0114	01109
Xml2	01	002	013	0004	0015	0114	

FIGURE 4.8 – tableau d'etiquetage final pour l'acp

Pour appliquer la régression il faut effectuer la transformation suivante : soit $\alpha \in etiquetage < xml > \in R^*$ $\alpha = (\alpha_1, \alpha_2, ..., \alpha_n)$ $\varphi' : etiquetage \rightharpoonup R$

$$\varphi'(\alpha) = \alpha_1 \times 1 + \alpha_2 \times 2 + ... + \alpha_n \times n$$

on obtient comme resultat la figure 4.9 :

Xml1	1	18	9	24	18	44	135
Xml2	1	6	9	24	45	44	

FIGURE 4.9 – tableau d'etiquetage final pour la regression

Pour l'ACP on calcule le premier vecteur propre qui servira de representation comporte des documents xml, selon le choix de l'orientation de la matrice, la taille peut etre soit la profondeur de l'arbre soit le nombre des éléments du document, qui sont tous les deux lineaires
Quant a la regression on choisi deux fonctions simples :
Lineaire $\rightarrow ax + b$ non lineaire $\rightarrow$ (parabole) $ax2 + bx + c$

4.4.2 Cas 2 : Entrepôt de données

Définition

Selon Bill Inmon, connu comme étant le père du Data Warehouse,"un Entrepôt de données est une collection de données thématiques, intégrées, non volatiles et historisées pour la prise de décisions".

De manière plus concrète, nous pouvons le définir comme une structure pour l'organisation des systèmes d'information. Il s'agit un processus d'aide à la prise de décision et la gestion de la connaissance tant pour l'usage quotidien que pour l'élaboration de stratégies à long terme.[7]
La figure 4.10 offre une vision générale de l'architecture d'un entrepôt de données.
Parmi les composantes de cette architecture, on distingue :

1. Sources de données ...
2. Back-end tier ...
3. Data Warehouse tier ...
4. OLAP tier ...
5. Front-end tier ...

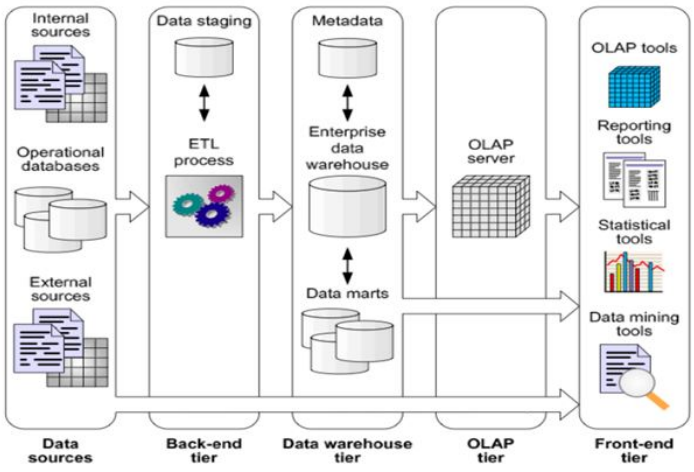

FIGURE 4.10 – architecture d'un entrepôt de données

Un entrepôt de données est basé sur un modele multidimensionnel où les données sont vues comme des data cubes **" cube de données"**

(Hyper) cube de données

Un cube represente un ensemble de mesures organisées selon un ensemble de dimensions. Une dimension est un axe d'analyse c'est-a-dire une base sur laquelle seront analysées les données(voir figure 4.11.

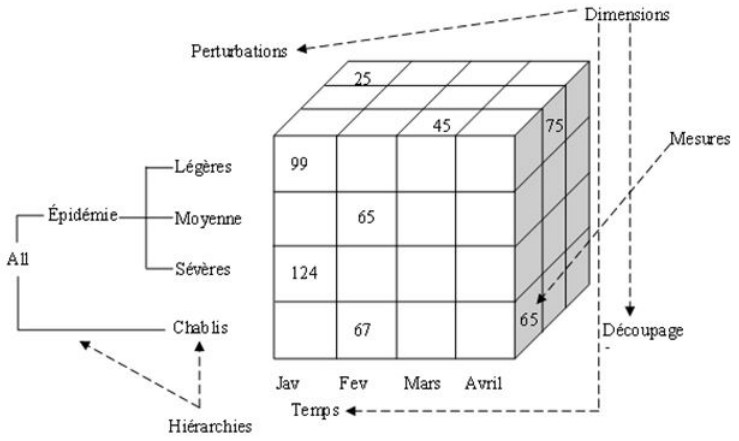

FIGURE 4.11 – exemple de cube de donnés

On-Line Analytical Processing (OLAP)

OLAP est l'ensemble des technologies qui, se basant sur une representation multidimensionnelle des données, permet aux analystes et decideurs de traiter leurs données de façon analytique, interactive (sessions), rapide et permettant de voir les données de l'entreprise sous plusieurs angles (dimensions).[21]

OLAP et les entrepôts de données sont complémentaires. Un entrepôt de donnees stocke et gére les données. OLAP transforme les donnees de l'entrepot en informations strategiques. OLAP peut passer d'une navigation basique a des calculs ou des analyses plus serieuses comme les series temporelles ou la modelisation complexe. Ainsi, les décideurs expérimenteront les capacités avancées d'OLAP, et pourront passer d'accés aux données a information, a connaissance. ".

3 grands types de stratégies d'implementation d'un entrepot de données et d'analyse possibles dans les produits OLAP :

ROLAP (Relational OLAP)

ROLAP Comme son nom l'indique, il utilise le concept relationnel pour stocker des données modelisées dans le format multi dimensionnel. Les analyses (drill-down, pivot, ajout de dimensions, etc.) sont transformees en requêtes SQL classiques qui sont executées sur les tables. R-OLAP utilise aussi la notion de tables d'agregats, c'est-a-dire créer des tables contenant des données sommaires et les stocker en mémoire en cas d'utilisation.[21]
la figure 4.12 montre un exemple de moteur ROLAP : Mondrian

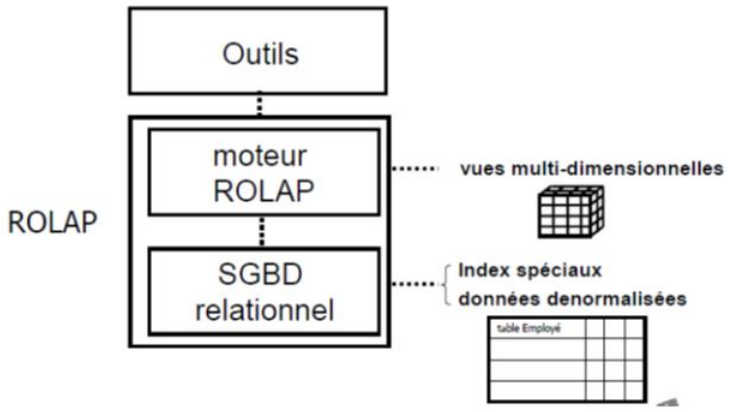

FIGURE 4.12 – le moteur ROLAP : Mondrian

ROLAP = Base de données relationnelle + SQL avancé

MOLAP (Multi dimensional OLAP)

Contrairement a R-OLAP, M-OLAP permet de stocker les données directement en un format permettant des opérations matricielles. Selon le constructeur, on trouvera un mode de stockage à base de tableau de données, de technologies propriétaires et même à base de fichiers plats.L'avantage de ce mode de stockage est la capacité a effectuer des calculs trés poussés en un temps record vu que tous les calculs sont précompilés.Le mode de stockage permet de pré calculer les resultats afin d'avoir accés directement a toute donnée, quel que soit le niveau de détail.[21]
la figure 4.13 represente un exemple de moteurs MOLAP : Microsoft Analysis Services, Hyperion

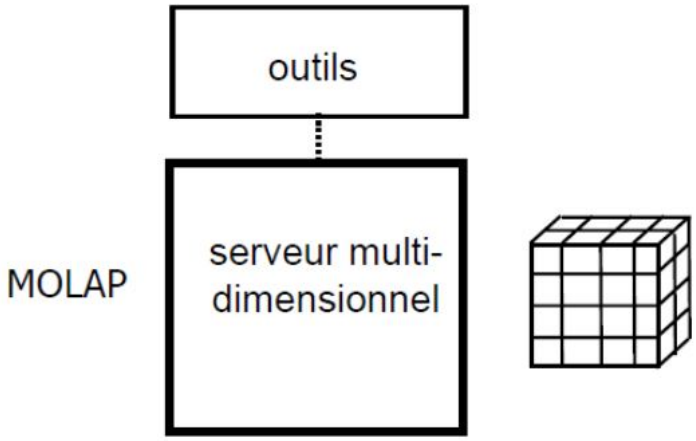

FIGURE 4.13 – moteur MOLAP : Microsoft Analysis Services, Hyperion

MOLAP = Base de données dimensionnelle + Serveur de traitement OLAP

HOLAP (Hybrid OLAP)

C'est la solution " en vogue " du moment, car elle permet de minimiser les défaillances des technologies R-OLAP et M-OLAP. Il s'agit en fait d'un mix des deux solutions.[21]
On utilisera un mode de stockage propriètaire pour les tables d'agrégat et les tables intermédiaires (permettant de ne pas avoir les points faibles du RO-LAP) On conservera un mode relationnel pour les tables de bas niveau.(voir figure4.14

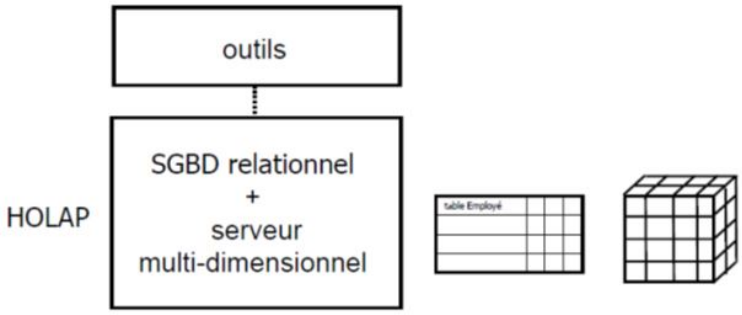

FIGURE 4.14 – exemple HOLAP

4.5 Implementaion

on a opté pour MATLAB pour implementer les deux techniques vu sa simplicité et sa puissance dans le calcul matriciel complexe.
et en plus la fonction de l'ACP et de la regression sont prédifinies.mais une petite partie du programme est implementée sous le langage C car on a rencontré quelque difficulté sous MATLAB
vu la contrainte de temps,nous avons juste implementé la premiére étude de cas :"document xml".

4.5.1 pseudo-code du programme

premierement on prend en entrée un document xml.(voir figure4.15

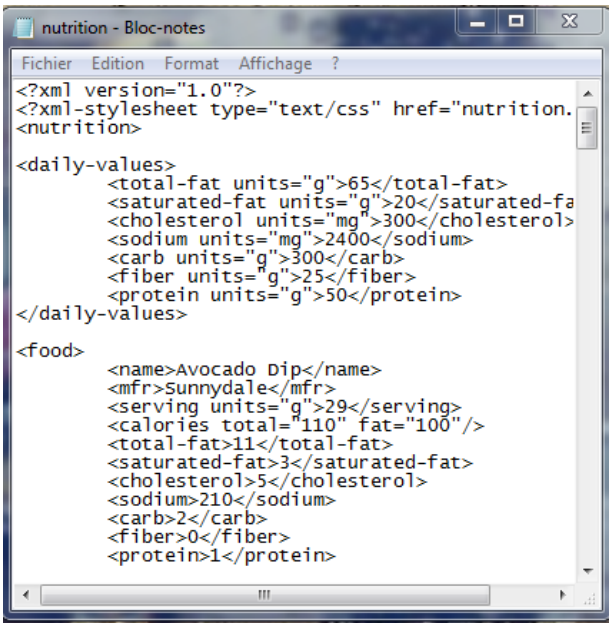

FIGURE 4.15 – document xml

doit transformer ce documenten arbre DOM(voir figure4.16

```c
xmlDocPtr doc;
xmlNodePtr racine;

// Ouverture du fichier XML
doc = xmlParseFile("nutrition.xml");
if (doc == NULL) {
    fprintf(stderr, "Document XML invalide\n");
    return EXIT_FAILURE;
}
// Récupération de la racine
racine = xmlDocGetRootElement(doc);
if (racine == NULL) {
    fprintf(stderr, "Document XML vierge\n");
    xmlFreeDoc(doc);
    return EXIT_FAILURE;
}
printf("La racine du document est : %s\n", racine->name);
// Libération de la mémoire
xmlFreeDoc(doc);

return EXIT_SUCCESS;
```

FIGURE 4.16 – transformation du document xml en arbre DOM

cette fonction est prédifinie en C.elle prend le nom de fichier,elle recupére la racine et puis elle libere la mémoire.
en second lieu,nous proposons la fonction d'étiquetage suivante :

```c
// fonction d'etiquetage
int etiq(noeud *n, int a) {

   i_root=1;

  for(j=1; j < n; j++)
  {
    i_noeud = i_father.n;

    for(i=1; (j < |cildren|; i++)
    {
      etiq(n.children_i.i)
      a[j]=n.etiq


  }
}
}
```

FIGURE 4.17 – fonction d'étiquetage

apré avoir étiqueter l'arbre DOM.nous trions le tableau résultant un tri de chaine de caractéres,et puis affecter a chaque donnée l'indice qui convient,comme suit 4.18 :

```
void trier(char tab[72][20], char mat[72][8]){
    int i,j;
    char tamp[20]={0}; char tamps[8]={0};
    for(i = 0; i < 72; i++)
    {
        for(j = i+1; j < 72; j++)
            if(strcmp(tab[i],tab[j]) > 0)  //Edit.
            {
                // permutation des données
                strcpy(tamp , tab[i]);
                strcpy(tab[i] ,tab[j]);
                strcpy(tab[j] ,tamp);
                // permutation des indices
               strcpy(tamps,mat[i]);
               strcpy(mat[i],mat[j]);
               strcpy(mat[j] , tamps);
                // copier les indice
            }
    }
}
```

FIGURE 4.18 – fonction de tri

les resultats de tri seront sauvegarder dans un tableau

pour apliquer l'ACP on doit calculer son propre tableau,comme suit 4.19

```
void concatination(char mat[72][8]){
    int i,j; char t[2];
    for (i=0;i<72;i++){
        j=i+1;
        itoa(j,t,10);
        strcat(mat[i], t);
    }

}
```

FIGURE 4.19 – fonction de calcul du tableau ACP

ainsi que pour la regression4.20

```
void calcule(char tab[72][8], int t[72]){
    int i,j,x,s=0; char tam[4];
    for(i=0;i<72;i++) {s=0;
        for(j=0;j<4;j++) {
            tam[j]=tab[i][j];
            x=atoi(tam);
            s=s+(x*j);
}
            t[i]=s;
}

}
```

FIGURE 4.20 – fonction de calcul du tableau de regression

les resultats des deux tableaux seront sauvegarder dans deux fichiers textes differents4.21

```
void fichier1(char tab[72][8])
{
FILE* fichier = NULL;
   int i;

   fichier = fopen("resultat_1.txt", "w");

   if (fichier != NULL)
   {

       // On l'écrit dans le fichier
       for (i=0;i<72;i++)
       fprintf(fichier, "%s \n",tab[i]);
       fclose(fichier);
   }

}
```

FIGURE 4.21 – sauvgarder les resultats dans des fichiers textes

pour le MATLAB on a cette interface,qui contient des boutons et un espace pour dessiner les resultats de calculs4.22

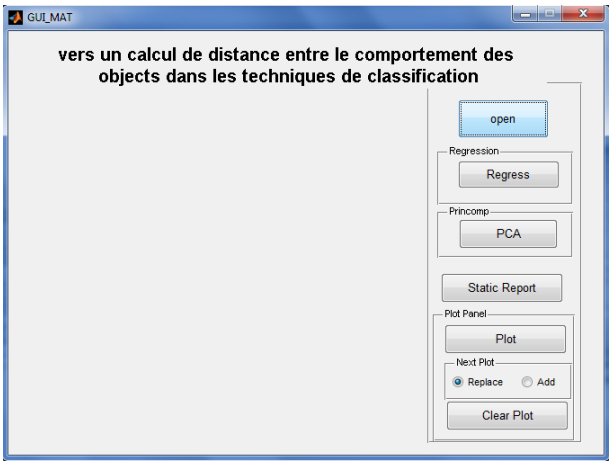

FIGURE 4.22 – interface du programme MATLAB

le bouton **open** ouvre le repertoire des fichiers4.23

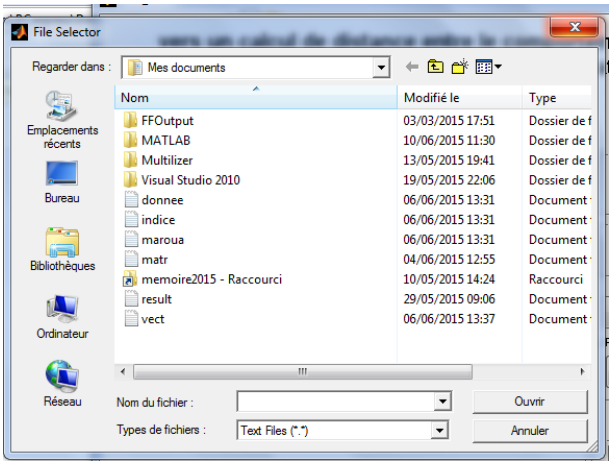

FIGURE 4.23 – repertoire des fichiers

le bouton **regress** fait un appel a la fonction de regression *regress* pour effectuer le calcul
le bouton **PCA** fait un appel a la fonction de l'ACP *princomp* pour le calcul

le bouton **static report** affiche le resultat dans un fichier texte4.24

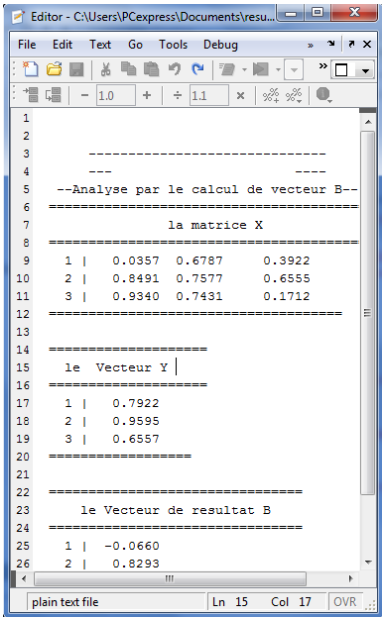

FIGURE 4.24 – resultat dans fichier texte

et le bouton **plot** affiche les resultats graphiques4.25

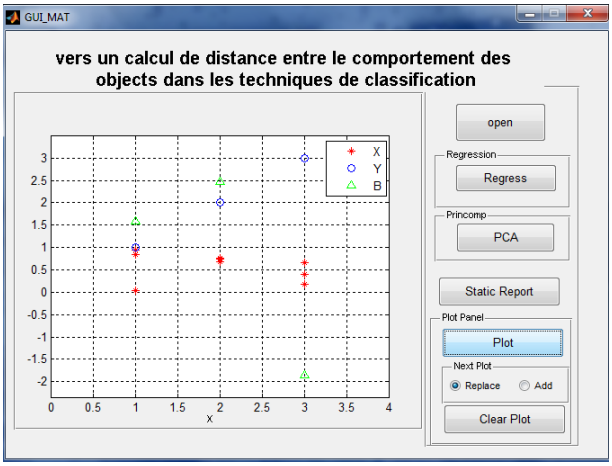

FIGURE 4.25 – resultat ploté

le bouton **clear plot** est pour supprimer les calculs et effectuer d'autres a nouveau

4.6 conclusion

dans ce chapitre nous avons fait une étude comparative entre les deux techniques et leurs implémentations...,selon quelques critéres

cette étude a été menné par une implementation des deux cas,
vu le facteur de temps nous avons pas pu implementer l'etude de cas des cubes de données

Chapitre 5

Conclusion générale

Dans ce mémoire, nous avons effectué une étude comparative entre la méthode de régression linéaire et non linéaire avec l'analyse en composantes principales dans le cadre du calcul de distance entre objets. Après avoir fixé un contexte et des critères de comparaison, nous avons montrer les points forts et faibles de chaque méthode ainsi que les parametres favorisant l'utilisation d'une methode vis a vis de l'autre.

Cette étude a été consolidée par une étude de cas, la premiere est une proposition d'un changement de représentation des documents XML. la deuxieme est la projection d'un cube de données vers un espace de dimension minimal.

Vu le temps qui nous a été alloué, seule la premiére étude de cas à été implémentée. Des difficultés comme le choix d'un cube de données peu complexe sont rencontrés.

Nous rappelons que cette implémentation ne fait pas de comparaison entre l'utilisation de ses deux méthodes et / ou l'une de ses méthodes avec le calcul de distance classique. Cet objectif laissé comme perspective exige la définition d'un corpus de donnée complet. d'effectuer un calcul de distance classique dans le cadre d'une tâche bien definie comme une classification par exemple. ensuite, l'application des deux techniques est appliquée. Une dernière phase de comparaison et de validation est demandée.

Bibliographie

[1] Algorithme espérance-maximisation sur le site de wikipedia.

[2] Pattern recognition and machine learning. 2006.

[3] S. E. J. e. M. Ali Kouani. Analyse en composantes principales une méthode factorielle pour traiter les données didactiques. 2007.

[4] A.Moussaoui. Classi cation dans le data mining. support de cours, ecole doctorale isi université de laghouat. 2006.

[5] A.Moussaoui. Classi cation dans le data mining. support de cours, ecole doctorale isi université de laghouat. 2006.

[6] A. Baccini and P. Besse. Data mining exploration statistique.technical report c5219, institut de mathématiques de toulouse, laboratoire de statistique et probabilités. 2007.

[7] M. A. Benitez-Guerrero E., C. Collet. « entrepôts de données : Synthèse et analyse », rapport de recherche. 1999.

[8] A. J. S. Bernhard Schölkopf. Learning with kernels : Support vector machines, regularization, optimization and beyond. 2002.

[9] s. l. c. l. . o. . Carole Albouy, « Il était une fois ... le data mining ».

[10] R. A. L. R. A. L. Carter. "the asymptotic distribution of fix-point least squares" , publié par dept. of economics, university of western ontario. 1973.

[11] P. Casin. Analyse des données et des panels de données. de boeck university, première edition. 1999.

[12] d. d. g. d. l. c. Claude Bruxelle. Construction et dimensionnement de la chaussée. 2007.

[13] consulter le site. http ://www.liafa.jussieu.fr/ carton/enseignement/xml/cours/program mation/. date de cosultaion :25 fevrier 2015.

[14] A. des Sciences. Rapport sur la science et la technologie n8, la statistique. 2000.

[15] L. Esch. Mathématique pour économistes et gestionnaires. 2007.

[16] L. Esch. Mathématique pour économistes et gestionnaires, volume 1.de boeck université, 3 ème edition,. 2007.

[17] N. P. et Lakhmi Jain. Advanced techniques in knowledge discovery and data mining. 2005.

[18] J. Han and M. Kamber. Data mining, concepts and technics.morgan kaufman. 2006.

[19] D. T. s.-i. Isabel Bloch. Filtrage d'images. 2007.

[20] N. C. John Shawe-Taylor. Support vector machines and other kernel-based learning methods. 2000.

[21] M. R. Kimball R. Entrepots de donnees : guide pratique de modélisation dimensionnelle, ed. vuibert. 2003.

[22] le site de l'Université Paris-Est Marne-la Vallée(UPEM). Le data mining.

[23] T. M. Mitchell. Machine learning. 1997.

[24] A. B. T. Morineau A. L'analyse en composantes principales. cisia, paris. 1998.

[25] philippe choquette. nouveaux algorithmes d'apprentissage pour les classifieurs de type scm. 2007.

[26] M. Sebag. Apprentissage non supervisé et satisfaction des contraintes.support de cours. 2005.

[27] T. Tim Bray and C. M. S.-M.-F. Y. Netscape, Jean Paoli. Extensible markup language (xml) 1.0. 2008.