1. **Introduction**

De nos jours, la qualité des données stockées augmente d’une manière exponentielle. Pour cela, l’extraction des connaissances devient une tache de plus en plus difficile. Si l’information visuelle est manipulée d’une manière involontaire et spontanée par le cerveau humain dans la représentation visuelle des données et des modèles de connaissances extrait du Data Mining demeure indispensable.

Elle facilite la reconnaissance des différents points d’intérêt, la génération d’hypothèse et la direction de la réflexion de l’utilisateur.

En effet le Data Mining, connu aussi sous le nom de « fouille de données » est un processus d’analyse de données qui cherche à extraire des connaissances à partir d’une grande quantité de données stockées dans des bases ou entrepôts de données hétérogènes et peu structurés .

Aujourd’hui le Data Mining est utilisé par plusieurs entreprises. Il permet à celle-ci de déterminer les liens entre les facteurs internes tel que, le positionnement du produit ou les compétences du personnel, et les facteurs externes tels que les indicateurs économiques, la concurrence, la démographie.

Ainsi le Data Mining est particulièrement utile dans plusieurs domaines d’activités.

**I.1 Domaines d’application de Data Mining**

Dans toutes les domaines d’application de Data Mining, on peut citer par exemple :

1. Le domaine médicale et pharmaceutique :
2. Les banques et les finances :
3. Vente, distribution et marketing :
4. Grande distribution :
5. Web Mining :
6. Text Mining :
7. Scientifique :
8. Sécurité informatique :
9. Environnement :
10. Santé :
11. Biologie :

**I.2 Evolution de Data Mining**

Préhistoire

1875 : régression linéaire de Francis Galton

1895 : formule du coefficient de corrélation de Karl Pearson

1900 : distribution de x2 de Karl Pearson

1936 : analyse discriminante de Fisher et Mahalanobis

1941 : analyse factorielle des correspondances de Gutmann

1943 : réseaux de neurones de McCulloch et PIHS

1944 : régression logistique de Joseph Berkson

1958 : perception de Rosenblatt

1962 : analyse des correspondances de J. P. Benzécri

1964 : arbre de décision AID de J. P. Sonquistet et J. A. Morgan

1965 : méthode des centres mobiles de E. W. Forgy

1967 : méthode de k-means de Mac Queen

1972 : modèle linéaire généralisé de Nelder et Wedderburn.

**Histoire**

1975 : algorithmes génétiques de Holland

1975 : méthode de classement DISQUAL de Gilbert Saprota

1980 : régression PLS de Herman et Svante Wold

1984 : arbre CART de Breiman, Friedman, Olshen, Stone

1986: perceptron multicouches de Rumenlhart et McClelland

1989 : réseaux de T. Kohonen (cartes auto-adaptives)

Vers 1990 : apparition du concept de Data Mining

1991 : méthode MARS de Jérôme H. Friedman

1993 : arbre C4.5. de J Ross Quinlan

1996 : bagging (Breiman) et boosting (Freund-Shapire)

1998 : support vector machines de Vladmir Vapnik

2001 : forets aléatoires de L. Breiman

2005 : méthode elasticnet de Zhou et Hastie

**I.3 Intérêt de Data Mining**

Les entreprises sont inondées de données (scanners des super marchés, internet, bases de données, entrepôts de données, etc…). Le Data Mining permet d’exploiter ces données pour améliorer la rentabilité d’une activité.

Le Data Mining permet d’augmenter le retour dur l’investissement d’un système d’information.

**I.4 Finalités : comprendre et décider, savoir, et prévoi**r

Le Data Mining est un outil qui permet de produire des connaissances :

* Dans le but de comprendre le phénomène dans un premier temps : Savoir
* Dans le but de prendre des décisions dans un second temps : Prévoir pour Décider

**I.5 méthodes du data Mining**

* Du bon sens : Il s’agit d’abord d’analyser les données avec du bon sens et un bon outil mathématique et statistique élémentaire.
* Des algorithmes, des calculs statistique : il s’agit d’appliquer des algorithmes de calcule des données. Ces algorithmes sont plus ou moins complexes à mettre en œuvre. Ils permettent de classer des valeurs et

I.6 Les techniques du Data Mining

La production des règles de raisonnement se fait à partir de plusieurs techniques plus ou moins spécifiques au Data0 Mining. Ces techniques combinent à la fois les statistiques et l’algorithmique. Globalement, on peut dire que certaines techniques visent à classer d’autres visent à prédire.

Les 6 grands types de technique de Data Mining sont :

1. La description (technique descriptive)

Principes

La description consiste à mettre à jour :

* Pour une variable donnée : la répartition de ses valeurs (tri, histogramme, moyenne, maximum, minimum, etc…)
* Pour deux ou trois variables données : les liens entre les répartitions des valeurs des variables. Ces liens s’appellent des « tendances ».

Intérêt

Favoriser la connaissance et la compréhension des données.

Méthodes

Méthodes graphiques pour la clarté : analyse exploratoire des données.

Exemples :

Répartition des achats par âge (lien entre les variables « achat » et « âge »).

1. La classification (technique descriptive)

Principe

La classification consiste à créer des classes (c.-à-d. des sous-ensembles) de données similaires entre elles et différentes données d’une autres classe(l’intersection des classes entre elles doit toujours être vide).

Intérêt

* Favoriser la compréhension et la prédication.
* Fixer les segments qui serviront d’ensemble pour des analyses approfondies.
* Réduire les dimensions : le nombre d’attributs quand il y en a trop au départ.

Méthodes :

* Classification hiérarchique
* Classification des k moyennes
* Réseaux de Kohonen
* Règles d’association

Exemples :

* Métatypologie d’une clientèle en fonction de l’âge, les revenus, le caractère urbain ou rural, la taille des villes, ….
* Pour un audit comptable, classer un comportement financière en catégorie normale ou suspecte.

1. L’association

Principe :

L’association consiste à trouver quelles valeurs des variables vont ensemble. Par exemple telle valeur d’une variable avec telle autre valeur de la variable.

Les règles d’association sont de la forme : si antécédent, alors conséquence. L’association ne fixe pas de variable cible. Toutes les variables peuvent à la fois jouer le rôle de prédicateurs et de variable cible. On appelle aussi ce type d’analyse une « analyse d’affinité »

Intérêt

Mieux connaitre le comportements.

Méthodes :

* Algorithmes a priori
* Algorithmes du GRI (introduction de règles généralisée)

Exemples :

* Analyse du panier de la ménagère (si j’achète du riz, alors j’acheté du haricot)
* Etudier quelle configuration contractuelle d’une abonne d’une compagnie de téléphone mobile conduit plus facilement a un changement d’opérateur.

1. Estimation (technique prédicative)

Principe :

L’estimation consiste à définir le lien entre un ensemble de variables indépendantes et une variable cible.

Ce lien est défini à partir des données complètes c.-à-d. dont les valeurs sont connues tant pour les prédicateurs que pour la variable cible. Ensuite, on peut déduire une variable cible inconnue de la connaissance des prédicateurs.

A la différence de la segmentation qui travaille sur une variable catégorielle, l’estimation travaille sur une variable cible numérique.

Intérêt :

Permettre l’estimation des valeurs inconnues.

Méthodes :

* Analyse statistique classique : régression linéaire simple, corrélation, régression multiple, intervalle de confiance, estimation de points.
* réseaux de neurones.

Exemples

* Estimer la pression sanguine à partir de l’âge, sexe, poids et le niveau de sodium dans le sang.
* Estimer les résultats dans les études supérieures en fonction des critères sociaux.

1. Segmentation (technique prédicative)

Principe :

La segmentation est une estimation qui travaille sur une variable cible catégorielle. On parle de segmentation car chaque valeur possible pour la variable va devoir définir un segment (ou type, ou classe, ou catégorie) de donnée.

La segmentation peut être vue comme une classification supervisée.

Intérêt :

Permettre l’estimation des valeurs inconnues

Méthodes :

* Graphiques et nuage des points
* Méthodes des k-plus proches voisins
* Arbre de décision
* Réseaux de neurones

Exemples :

* Segmentation par tranche de revenus : élevé, moyen, faible (3 segments). On cherche les caractéristiques qui conduisent à ces segments
* Déterminer si un mode de remboursement présente un bon ou un mauvais risque crédit (2 segments).

1. La prévision (technique prédicative)

Principe :

La prévision est similaire à l’estimation et la segmentation mise à part que pour la prévision, les résultats portent sur le futur.

Intérêt :

Permettre l’estimation des valeurs inconnues.

Méthodes :

Celles de l’estimation et de la segmentation.

Exemple :

* Prévoir le prix a 3 mois dans le futur.
* Prévoir le temps qu’il va faire.
* Prévoir le gagnant du championnat de football, par rapport à une comparaison des résultats des équipes.

I.7 Les logiciels utilisés en Data Mining

Les logiciels qu’on peut utiliser en Data Mining sont nombreux et diversifier, on peut citer, par exemple :

1. Logiciels usuels de statistique : SAS, R, SPlus, …
2. Logiciels de calcul numérique : Matlab, MS Excel, …
3. Parties des logiciels dédiés : SAS/EM, SPSS/Clementine, statsoft/statistica Data Miner, Insight/Insightful Miner, ….
4. Les logiciels libres : Weka, tanagra, Galicia, …
5. Des systèmes de gestion des bases de données : Access, SQLServer, Oracle, …
6. Analyse descriptive d’une variable

* Réduire et résumer les données

On peut distinguer plusieurs types de paramètre selon l’objet :

* Paramètres de position ou de tendance centrale.
* Paramètres de position.
* Paramètres de dispersion.
* Paramètres de symétrie.
* Paramètres de forme

II.1 paramètres de tendance centrale ou de centralité

On distingue 3 paramètres centraux ou de tendance centrale : moyenne, médiane et mode.

II.1.1 Moyenne : ordre de grandeur de données

Définition : la moyenne est la valeur obtenue lorsque les valeurs sont égales, lorsque la distribution est équitable

II.1.1.1 Moyenne arithmétique

La moyenne arithmétique est la somme des valeurs et leurs nombres.

Moyenne d’une série classée

Elle concerne une série classée ou à chaque valeur correspond un effectif, les valeurs xi se répètent plus d’une fois dans une distribution. La moyenne est égale aux centres des classes (xi) pondérés par leurs effectifs correspondants (ni) qu’on rapporte un effectif global (n) :

Xa =

Le calcul de la moyenne passe par les 4 étapes suivantes :

1. Déterminer le centre des classes (xi)
2. Multiplier les centres des classes (xi) par l’effectif correspondant à la classe (ni)
3. Etablir la somme des produits :
4. Diviser cette somme par l’effectif global(n) :

Exemple : moyenne d’une série classée

Il s’agit de calculer la moyenne de la taille des exploitation d’un périmètre irrigue dont la structure est résumée par le tableau suivant :

|  |  |  |  |
| --- | --- | --- | --- |
| Taille en ha  X | Nombre exploitant  Ni | Centre des classes  Xi | Superficie  Si = |
| 0 – 5 | 400 | 2.5 | 1000 |
| 5 – 10 | 50 | 7.5 | 375 |
| 10 – 20 | 35 | 15 | 525 |
| 20 – 50 | 15 | 35 | 525 |
| Total | 500 |  | 2425 |

Xa =

II.1.1.2 La moyenne géométrique

La moyenne géométrique est la racine nième du produit des valeurs xi : Xg =

II.1.1.3 Moyenne harmonique

Elle est égale à l’inverse de la moyenne arithmétique des inverses

II.1.1.4 Moyenne quadratique : les surfaces

La moyenne quadratique est la racine carrée de la moyenne arithmétique des carrées des valeurs

II.1.1.5 Formule générale de la moyenne

II.1.2 médiane

1. Distribution impaire
2. Distribution paire

II.1..3 Le mode : Mo

C’est la valeur ou la classe la plus fréquente de la série.

II.2 Les paramètres de position : les quantiles

II.3 les paramètres de dispersion

1. Variance

C’est le moment centre de

1. L’écart-type

II.4 Standardisation : comparaison

La standardisation simplifie la lecture des données et permet de voir plus rapidement les liens entre les variables surtout lorsqu’on a à faire a plusieurs variables ou à un grand tableau. On peut voir très rapidement les différentes corrélations entres les variables.

Il suffit pour cela de voir les signes des valeurs standardisées. Les variables corréler positivement auront partout le même signe, celles qui varient en sens inverse auront des signes opposées. Enfin les variables qui sont peu corréler ont tendance à avoir une disposition aléatoire des signes.

1. Analyse de la relation entre deux variables

III.1. Corrélation et régression linéaires simples

Modèle :

1. Déterminez les valeurs des paramètres a et b
2. Par la méthode des moindres carrés.
3. En utilisant les formules de calcul de la covariance, covariance et écart-type.
4. Donnez le modèle sous la forme  ;  ; avec
5. Calculez le coefficient de corrélation
6. Calculer le coefficient de détermination
7. Déterminez la variance expliquée.
8. Déterminez la variance résiduelle.
9. Coefficient de corrélation

….

1. Méthode de calcul matriciel
2. Méthode général

Le modèle général s’écrit : avec Y : variable dépendante, X : variables explicatives et A : les coefficients de régression . La résolution de cette équation passe par l’inversion de la matrice X, qui représente les données brutes Dyx, les variances Vyx ou les corrélations simples Ryx :

Matrice des données

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Y | X1 | X2 | … | Xn |
| 1 | Y1 | X11 | X12 | … | X1n |
| 2 | Y2 | X21 | X22 | … | X2n |
| … | … | … | … | … | … |
| N | Yn | Xn1 | Xn2 | … | Xnn |

Lorsqu’on dispose des variances et des covariances on utiliser la matrice des variances-covariances pour calculer le coefficient de corrélation multiple :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Y | X1 | X2 | … | Xn |
| Y | Vy | COVyx1 | COVyx2 | … | COVyxn |
| X1 | COVyx1 | Vx1 | COVx1x2 | … | COVx1xn |
| X2 | COVyx2 | COVx1x2 | Vx2 |  | COVx2xn |
| … | … | … | … | … | … |
| Xn | COVxn | COVx1xn | COVx2xn | … | Vxn |

Lorsqu’on dispose des coefficients de corrélation simple, on peut utiliser la matrice des corrélations pour calculer le coefficient de corrélation multiple :

Matrice des corrélations

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Y | X1 | X2 | … | Xn |
| Y | 1 | Ryx1 | Ryx2 | … | Ryxn |
| X1 | Ryx1 | 1 | Rx1x2 | … | Rx1xn |
| X2 | Ryx2 | Rx1x2 | 1 |  | Rx2xn |
| … | … | … | … | … | … |
| Xn | Ryxn | Rx1xn | Rx2xn | … | 1 |

A partir de l’une de ces matrices on calcule la corrélation multiple, les coefficients de régression et la variance résiduelle.

1. La corrélation multiple

En utilisant la matrice des corrélations simples, on peut écrire la formule :

Δ11 = le niveau en supprimant la ie ligne et colonne.

1. Cas de 2 variables

Lorsqu’on a 2 variables seulement, on a le chemin suivant :

|  |  |  |  |
| --- | --- | --- | --- |
|  | Y | X1 | X2 |
| Y | 1 | Ryx1 | Ryx2 |
| X1 | Ryx1 | 1 | Rx1x2 |
| X2 | Ryx2 | Rx1x2 | 1 |

1. Cas de 3 variables

….

1. Cas de 4 variables
2. Régression

En utilisant le calcul matriciel, on a l’équation matriciel suivante :

Pour déterminer les paramètres , il suffit d’inverser la matrice X ; on obtient alors

et

y =

b =

y =

b =

1. **Cas des variances et covariances**

Si on dispose des variances et des covariances, on peut procéder comme suit :

D’où avec

[COVyx] :vecteur des covariances yx ;

[Vx] : matrice des variances et des covariances ;

[A] : vecteur des paramètres de régression

On a alors

|  |  |  |  |
| --- | --- | --- | --- |
|  | Y | X1 | X2 |
| Y | Vy | Cov(y,x1) | Cov(y,x2) |
| X1 | Cov(y,x1) | Vx1 | Cov(x1,x2) |
| X2 | Cov(y,x2) | Cov(x1,x2) | Vx2 |

Delta

Delta 1

Delta 2

**La matrice des données**

Y =

F =

F=

F=

=

=

=

Equation

**Le modèle linéaire**

Y = x = a = b =

**La droite des moindres rectangles**

Y =

Avec

Linéaire

2. La droite des moindres rectangles

La droite des moindres distances

Y =

A =

b =

a =

b =

**Modèle polynomiale**

y = ;

y = ;

y = ;

y = ;

y = ;

y =

;

y = ;

F =

Coefficient de détermination =

**Modèle exponentiel**

**Modèle puissance**

On a ⬄