République Algérienne Démocratique et Populaire

Ministère de l’Enseignement Supérieur et de la Recherche Scientifique

**Université des Sciences et de la Technologie Houari Boumediene**

Département Informatique

# Module : Data Mining

**Résumé du livre « Data Mining concepts and techniques »**

**Réalisation par :**

**-CHABANE Nouar 201400007379**

**-BEZGALI Meriem 201300005476**

**Groupe : 01**

**Chapitre 01 :**

De façon analogue, l’exploration de données aurait dû être nommée «exploration de connaissances à partir de données», ce qui est malheureusement quelque peu long à prononcer.

Beaucoup de gens considèrent l’exploration de données comme un synonyme d’un autre terme couramment utilisé, « découverte de connaissances à partir de données », tandis que d’autres considèrent l’exploration de données comme une simple étape essentielle dans le processus de découverte des connaissances. Le processus de découverte des connaissances est constitué des étapes suivantes :

* Nettoyage des données (pour éliminer le bruit et les données incohérentes)
* Intégration des données (lorsque plusieurs sources de données peuvent être combinées)
* Sélection des données (où les données pertinentes pour la tâche d'analyse sont extraites de la base de données)
* Transformation de données (où les données sont transformées et consolidées dans des formulaires approprié pour l'extraction en effectuant des opérations de résumé ou d'agrégation)
* Data mining (un processus essentiel où des méthodes intelligentes sont appliquées pour extraire un modèle de données)
* Évaluation des modèles (pour identifier les modèles vraiment intéressants représentant la connaissance sur la base de mesures d’intérêt)
* Présentation des connaissances (où techniques de visualisation et de représentation des connaissances sont utilisées pour présenter aux utilisateurs les connaissances acquises)

Les étapes 1 à 4 sont différentes formes de prétraitement des données, dans lesquelles les données sont préparées pour le Data Mining. L'étape d'exploration de données peut interagir avec l'utilisateur ou une base de connaissances.

Les modèles intéressants sont présentés à l'utilisateur et peuvent être stockés en tant que nouvelles connaissances dans la base de connaissances.

En tant que technologie générale, l'exploration de données peut être appliquée à tout type de données à condition que les données aient un sens pour l'exploration.

Nous utilisons une base de données et elle est généralement décrite avec un modèle sémantique nommée entité-relation auquel on accède à l'aide de requêtes.

Outre les données de base de données relationnelles, «data warehouse » et les données de transaction, il existe de nombreux autres types de données qui ont des formes et des structures variées et des significations sémantiques assez différentes, nous avons observé différents types de référentiels de données et d'informations sur lesquels l'exploration de données peut être effectuée. Un certain nombre de fonctionnalités d’exploration de données sont utilisées pour spécifier les types de modèles que l’on veut trouver dans les tâches d’exploration de données. En général, ces tâches peuvent être classées en deux catégories: descriptive et prédictive.

Les données d’entrée peuvent être des classes ou des concepts. Les modèles qu’on veut souvent extraire sont des modèles qui nous aides a effectuer des classifications d’objets, leurs cluster ou l’analyse des valeurs aberrantes.

**1. Classification:**

Est le processus de recherche d’un modèle (ou d’une fonction) qui décrit et distingue les classes de données ou des concepts. Les modèles sont dérivés en fonction de l'analyse d'un ensemble de données d'apprentissage (c'est-à-dire des objets de données pour lesquels les étiquettes de classe sont connues). Le modèle est utilisé pour prédire l'étiquette de classe d'objets dont l'étiquette de classe est inconnue et présentée sous différentes formes.

**2. Clustering:**

Analyse les objets de données sans consulter les étiquettes de classe. Dans de nombreux cas, les étiquetés des données peuvent tout simplement ne pas exister au début pour un groupe de données. Les objets sont regroupés ou regroupés selon le principe de maximisation de la similarité intraclasse et de minimisation de la similarité interclasse à partir de chaque règle de cluster.

**3. Outlier Analysis:**

Un ensemble de données peut contenir des objets non conformes au comportement ou au modèle général des données. Ces objets de données sont des valeurs aberrantes. De nombreuses méthodes d'exploration de données rejettent les valeurs aberrantes comme du bruit ou des exceptions. Cependant, dans certaines applications (par exemple, la détection de fraude), les événements rares peuvent être plus intéressants que ceux qui se produisent plus régulièrement. L'analyse de les données sur les valeurs aberrantes sont appelées analyse des valeurs aberrantes ou extraction d'anomalies.

Tous les modèles ne sont pas tous intéressants, cela dépend de ce que l'utilisateur recherche et des caractéristiques principales qu'il recherche dans les modèles.

Dans cette section, nous donnons des exemples de plusieurs disciplines qui influencent fortement le développement de méthodes d’exploration de données.

**1. Statistique:**

Les statistiques étudient la collecte, l'analyse, l'interprétation ou l'explication et la présentation de données. L'exploration de données a un lien inhérent avec les statistiques. Un modèle statistique est un ensemble de fonctions mathématiques décrivant le comportement de les objets d'une classe cible en termes de variables aléatoires et leur probabilité associée distributions. Les modèles statistiques sont largement utilisés pour modéliser des données et des classes de données.

**2. Machine learning:**

L'apprentissage automatique étudie comment les ordinateurs peuvent apprendre (ou améliorer leurs performances) basé sur des données. C’est un domaine de recherche principal qui concerne les programmes informatiques pour apprendre automatiquement à reconnaître des modèles complexes et prendre des décisions intelligentes basées sur des données. Par exemple, un problème typique d’apprentissage automatique est de programmer un ordinateur pour qu’il puisse automatiquement reconnaître les codes postaux. L'apprentissage automatique est une discipline en pleine croissance.

**3. Database systems research :**

La recherche sur les systèmes de bases de données porte sur la création, la maintenance et l'utilisation de bases de données pour les organisations et les utilisateurs. En particulier, les chercheurs en systèmes de bases de données ont établi des principes hautement reconnus dans les modèles de données, les langages de requête, le traitement des requêtes méthodes d’optimisation, stockage de données et méthodes d’indexation et d’accès de base de données. Ces systèmes sont souvent bien connus pour leur grande évolutivité dans le traitement de très gros ensembles de données.

**4. Recherche d’informations :**

La recherche d'information (IR) est la science de la recherche de documents ou d'informations dans les documents. Les documents peuvent être du texte ou du multimédia et peuvent résider sur le Web. Les différences entre la recherche d'informations traditionnelle et les systèmes de base de données sont doubles:

La recherche d'informations suppose que les données sous recherche sont non structurées; et les requêtes sont formées principalement par des mots-clés, qui n'ont pas de structures complexes (contrairement aux requêtes SQL dans les systèmes de base de données). Les approches typiques en matière de récupération d'informations adoptent des modèles probabilistes.

La principale application ciblée par l'exploration de données est l'intelligence d'affaires et les moteurs de recherche.

**Chapitre 02:**

Les ensembles de données sont constitués d'objets de données. Un objet de données représente une entité, exemple dans base de données de vente, les objets peuvent être des clients, des articles de magasin et des les ventes effectuée, dans une base de données médicale, le les objets peuvent être des patients; dans une base de données universitaire, les objets peuvent être des étudiants, des professeurs, et des cours. Les objets de données sont généralement décrits par des attributs. Les objets de données peuvent également être appelés échantillons, exemples, instances, points de données ou objets. Si les objets de données sont stockés dans une base de données, ils sont des tuples de données. C'est-à-dire que les lignes d'une base de données correspondent aux objets de données et les colonnes correspondent aux attributs.

Le type d'un attribut est déterminé par l'ensemble des valeurs nominales possibles, binaires, ordinal ou numérique (discret ou continu), notez que si un attribut n'est pas discret, il est continu. Les termes attribut numérique et attribut continu sont souvent utilisés de manière interchangeable dans la littérature.

* **Les statistiques et les données:**

Pour que le prétraitement des données réussisse, il est essentiel de disposer d’une image globale de votre système. Les données. Des descriptions statistiques de base peuvent être utilisées pour identifier les propriétés des données et mettre en évidence les valeurs de données à traiter comme du bruit ou des valeurs aberrantes, telles que la mesure de la tendance centrale: moyenne, médiane et mode, de gamme moyenne. Nous pouvons également connaître la dispersion des données: plage, en mesurant les quartiles, la variance, l’écart type et la plage interquartile nous pouvons utiliser une représentation graphique pour montrer des données plus clairement à l'utilisateur.

* **Matrice de données et matrice de dissimilarité:**

Matrice de données (ou structure objet par attribut): cette structure stocke les n objets de données sous la forme d'une table relationnelle ou d'une matrice n-par-p (n objets / p attributs): chaque ligne correspond à un objet. La matrice de dissemblance (ou structure objet par objet): cette structure stocke une collection des proximités disponibles pour toutes les paires de n objets. Il est souvent représenté par une table n-par-n, les mesures de similarité peuvent souvent être exprimées en fonction de mesures de dissimilarité.

Pour mesurer la dissimilarité entre les objets, nous calculons généralement la dissimilarité des objets décrits par des attributs numériques. Ces mesures comprennent les distances euclidiene, Manhattan et Minkowski.

**Chapitre 03:**

Il existe plusieurs techniques de prétraitement des données. Nous examinons les principales étapes du prétraitement des données, à savoir le nettoyage des données, l'intégration, la réduction et la transformation des données. Les routines de nettoyage des données permettent de «nettoyer» les données en remplissant les valeurs manquantes, en lissant les données bruitées, en identifiant ou en supprimant les valeurs aberrantes et en résolvant les incohérences. Si les utilisateurs croient que les données sont sales, il est peu probable qu'ils se fient aux résultats de l'exploration de données qui a été appliquée et cela peut créer une confusion.

1. **Data cleaning :**

Les méthodes du Data cleaning :

1. **Missing Values :**

* Ignorer le tuple: Ceci est généralement fait lorsque l’étiquette de classe est manquante (en supposant que la tâche d'extraction implique la classification). Cette méthode n’est pas très efficace, à moins que le tuple contient plusieurs attributs avec des valeurs manquantes. Il est particulièrement pauvre lorsque le pourcentage des valeurs manquantes par attribut varie considérablement.. Ces données pourraient avoir été utiles à la tâche à accomplir.
* Remplir manuellement la valeur manquante: en général, cette approche prend du temps et peut ne pas être réalisable étant donné un grand ensemble de données avec de nombreuses valeurs manquantes.
* Utilisez une constante globale pour renseigner la valeur manquante: remplacez toutes les valeurs d'attribut manquantes. par la même constante, telle qu'une étiquette telle que «Inconnu» ou 1. Si les valeurs manquantes sont remplacé par, disons, «inconnu», le programme minier peut penser à tort que ils forment un concept intéressant, car ils ont tous une valeur commune: celle de "Inconnu". Par conséquent, bien que cette méthode soit simple, elle n'est pas infaillible.
* Utilisez une mesure de la tendance centrale de l’attribut (par exemple, la moyenne ou la médiane) pour compléter la valeur manquante.
* Utilisez l'attribut moyenne ou médiane pour tous les échantillons appartenant à la même classe que le tuple donné.
* Utilisez la valeur la plus probable pour renseigner la valeur manquante: cela peut être déterminé avec régression, outils basés sur l'inférence utilisant un formalisme bayésien ou un arbre de décision

1. **Noisy Data :**

Le bruit est une erreur ou une variance aléatoire dans une variable mesurée. Étant donné un attribut numérique tel que, par exemple, le prix, comment pouvons-nous “Lisser” les données pour éliminer le bruit? Regardons le lissage des données suivant techniques. « Binning »: les méthodes de Binning lissent une valeur de données triée en consultant son "voisinage" c'est-à-dire les valeurs qui l'entourent et ils effectuent un lissage local.

**2) data integration :**

L’intégration de données différentes provoque des incohérences, une intégration minutieuse peut aider à réduire et à éviter les redondances et les incohérences.

**3) Data reduction :**

Pour obtient une représentation réduite de l'ensemble de données qui est beaucoup plus petite dans volume, tout en produisant les mêmes (ou presque) résultats analytiques. Les stratégies de réduction des données incluent la réduction de la dimensionnalité et la réduction de la numérosité. Des techniques de réduction des données peuvent être appliquées pour obtenir une représentation réduite du l’ensemble de données beaucoup plus petit en volume, tout en maintenant de près l'intégrité du document d'origine, c’est-à-dire que l’extraction de données réduites devrait être plus efficace tout en produisant la résultats d'analyse identiques (ou presque identiques).

**4) Data Transformation :**

C’est les méthodes de transformation de données. Dans cette étape de prétraitement, les données sont transformées ou consolidées afin que le processus d’extraction qui en résulte soit plus simple et efficace, et les modèles trouvés puisse être plus facile à comprendre.