Fouille de données

Mokeddem Sid Ahmed

Université de Mostaganem Abdelhamid Ibn Badis

Résumé Les algorithmes de fouille de données forment la base d'un domaine très à la mode au monde de traitement automatique des données c'est la "Science de données, ou bien Data Science" , ce domaine inclut des méthodes automatiques pour l'analyse des motifs et modèles pour tout type de données, ces applications ont une grande ampleur sur différents domaines tels que le business intelligence.

Ces notes de cours est destinées aux étudiants de Master 2 ISI. Elles permettent d'initier les étudiants au domaine de la science des données en présentant l'approche fouille de données en intégrant les concepts prérequis de l'apprentissage machine et statistique. Cependant, le document est structuré comme suivant :

— Analyse de données exploratoire
— Extraction des motifs fréquents (pattern mining)
— Clustering
— Classification

Ce document montre les bases de ces approches et couvre l'analyse en big data, avec des exemples et une comparaison des différents algorithmes, ce document est un guide de fouille de données pour les étudiants et les chercheurs.

Table des matières

Introduction

Ces dernières années, le domaine de science de données a connu une forte accélération avec l'apparition du phénomène BIG DATA.

Les caractéristiques des « données » ont singulièrement évolué surtout par l'évolution technologique (internet est un média incontournable, les capacités de stockages évoluent fortement, etc.), de par nos pratiques de communication (les réseaux sociaux, les forums, etc.), par la multiplication des sources et des formats des informations transmises (textes, images, vidéo avec les plates-formes d'échange, etc.). Ce phénomène désigne les caractéristiques des données par les termes : volume, variété, vélocité.

De ce fait, de nouveaux contraintes apparaissent pour soulever l'évolution de la spécialité où l'objectif reste néanmoins toujours la valorisation des données par des techniques informatiques tels que l'apprentissage machine ou statistique : big analytics, big data analytics, business analytics, etc. Ces techniques peuvent être rassembler sous le terme générique DATA SCIENCE, avec un nouveau métier : data scientist. Par conséquent, on peut dire que la fouille de données représente le noyau du data science d'où la nécessité de comprendre les fondement de ce domaine.
la fouille de données, dans sa forme et compréhension actuelle, à la fois comme champ scientifique et industriel, est apparu au début des années 90. Cette émergence n'est pas le fruit du hasard mais le résultat de la combinaison de nombreux facteurs à la fois technologiques, économiques et même sociopolitiques.

On peut voir la fouille de données comme une nécessité imposée par le besoin des entreprises afin de valoriser les données stockés dans leurs base de données. En effet, avec cette explosion en matière de stockage ce qui devient un enjeu pour les entreprises (la prise de conscience de l'intérêt commercial pour l'optimisation des processus de fabrication, vente, gestion, logistique, ...), beaucoup de questions s'imposent ou la plus principale est : Que doit-on faire avec des données coûteuses à collecter et à conserver ?
La fouille de données a aujourd'hui une grande importance économique du fait qu'elle permet d'optimiser la gestion des ressources (humaines et matérielle). Elle est utilisée par exemple :

- **Organisme de crédit :** pour décider d'accorder ou non un crédit en fonction du profil du demandeur de crédit, de sa demande, et des expériences passées de prêts ;
- **organisation des rayonnages dans les supermarchés** regroupant les produits qui sont généralement achetées ensemble (pour que les clients n'oublient pas bêtement d'acheter un produit parce qu'il est situé a l'autre bout du magasin). Par exemple, on extraira une règle du genre : " les clients qui achètent le produit X en fin de semaine, pendant l'été, achètent généralement également le produit Y " ;
- **Organisation de campagne de publicité**, promotions, ... (ciblage des offres)

— **diagnostic médical** : "les patients ayant tels et tels symptômes et demeurant dans des agglomérations de plus de 10 4 habitants développent couramment telle pathologie " ;
— **commerce électronique**, recommandation de produits
— **analyser les pratiques et stratégies commerciales** et leurs impacts sur les ventes
— **moteur de recherche sur internet** : *fouille du web*
— **extraction d'information depuis des textes** : *fouille de textes*

Dans ce cours, on s'intéressera essentiellement au différentes approches liées à la fouille de données avec l'étude de quelques exemples typiques de ces algorithmes constitue le corps de ce cours, suivie de l étude de quelques applications réelles. Avant tout, nous discutons de la notion de données.

1 Qu'est ce qu'une données ?

Cette section a pour objet de fixer un vocabulaire et de rappeler quelques faits importants concernant les attributs des données et ce que représente la valeur d'un attribut. Mais tout d'abord quelques notations que nous retrouverons dans l'ensemble du cours.

1.1 Notations

On notera D un ensemble de données, Chaque données est décrite par un ensemble de descripteurs (attributs) A. chaque attribut $a \in A$ prend sa valeur dans un certain nombre de valeurs V_a. Si on a p attributs $a_1, ..., a_p$. Ainsi, on peut considérer l'espace des données $E = V_{a_1} \times V_{a_2} \times ... \times V_{a_p}$ qui balayent toutes valeurs possible des attributs. Toute donnée appartient à cet ensemble de données $D \subset E$.
Il est utile d'avoir une représentation géométrique de l'espace des données E de P dimensions ou chaque attribut correspond à un axe.

1.2 Types d'attributs

Une donnée est un *enregistrement (tuple)* au sens des bases de données, que l'on nomme aussi "*individu*" (terminologie issue des statistiques) ou "*instance*" (terminologie orientée objet en informatique) et "*point*" ou "*vecteur*" parce que finalement, d'un point de vue abstrait, une donnée est un point dans un espace euclidien ou un vecteur dans un espace vectoriel. Une donnée d est décrite par un ensemble d'attributs d'attributs A. Un attribut a peut être de nature qualitative ou quantitative en fonction de V_a.
Attribut qualitatif, si on peut pas faire une moyenne (une couleur, une marque de voiture, ...). Sinon **l'attribut quantitatif** : un entier, un réel, ... ; il peut représenter un salaire, age, nombre d'habitants, etc., donc les opérations arithmétique habituels sont applicable ce qui est le cas des attributs qualitatifs.

Remarque : un attribut quantitatif ne signifie pas *numérique*, et réciproquement : un code postal est numérique mais pas quantitatif.

1.3 Type d'attributs par nature de leur valeurs

Si on veut un typage raffiné des attributs il faut analyser les valeurs de ces derniers. Naturellement ces valeurs sont censées représenter une certaine mesure afin de calculer une quantité dans ce monde. On voit donc apparaître des distinctions plus subtiles entre des attributs dont les valeurs sont arbitraires et incomparables **attribut nominal**, par exemple couleur, on peut pas comparer vert est blanc. On peut dire que la température aujourd'hui est 10 C et qu'hier, il faisait 18 C, on peut dire que la température était élevé hier par rapport à aujourd'hui, donc on peut comparer les valeurs, d'où un attribut, avec des valeurs arbitraires et incomparables, est un **attribut ordinal**. Les attributs avec des valeurs non arbitraires sont considérés comme **attribut absolu** , par exemple, le nombre d'enfant. Ces différentes natures entraînent le fait que les opérations que l'on peut faire sur ces attributs ne sont pas les mêmes.

1.4 Bruit

Un ensemble de données D est généralement contraint à des attributs avec des valeurs inconnue ou encore moins des valeur non valide ; il faut donc gérer des données dont certains attributs ont une valeur inconnue ou invalide ; on dit que les données sont **Bruitées**. La simple élimination des données ayant un attribut dont la valeur est inconnue ou invalide pourrait vider complètement la base de données ! alors quand on fait de la fouille de données on effectue de nombreuses opérations sur les données hors, selon la nature de l'attribut, ces opérations sont licites ou non... Il importe donc de ne pas faire n'importe quel calcul, d'appliquer n'importe quel algorithme sans prendre garde aux attributs sur lesquels on les effectue d'où la nécessite d'une phase de *prétraitement de données*.

2 Extraction des connaissances à partir des données

Une confusion subsiste encore entre fouille de données et knowledge discovery in data bases (KDD) (Extraction des connaissances à partir des données (ECD)). La fouille de données est l'un des maillons du processus de traitement pour la découverte des connaissances à partir des données. Sous forme imagée, nous pourrions dire que l'ECD est un véhicule dont la fouille de données est le moteur.
La fouille de données représente le noyau du processus de découverte des connaissances à partir des données. Les données peuvent être stockées dans des entrepôts (data warehouse), ou sur n'importe quel support de stockage. La fouille de données ne se limite pas au traitement des données structurées sous forme de tables ; il offre des moyens pour aborder les corpus en langage naturel(fouille de texte), les images (fouille des images), le son (sound mining) ou la vidéo et dans ce cas,

on parle alors plus généralement de multimedia mining.
L'ECD est un processus complexe qui se déroule suivant une série d'opérations. Des étapes de prétraitement ont lieu avant la fouille de données.

2.1 Étape de prétraitement

Le prétraitement porte sur l'accès aux données en vue de construire des datamarts, des corpus de données spécifiques. Le prétraitement concerne la mise en forme des données entrées selon leur type (numériques, symboliques, images, textes, sons), ainsi que le nettoyage des données, le traitement des données manquantes, la sélection d'attributs ou la sélection d'instances. Cette première phase est cruciale car du choix des descripteurs et de la connaissance précise de la population va dépendre la mise au point des modèles de prédiction. L'information nécessaire à la construction d'un bon modèle de prévision peut être disponible dans les données mais un choix inapproprié de variables ou d'échantillon d'apprentissage peut faire échouer l'opération.

Bibliographie

Bell, J. (1964). On the Einstein Podolsky Rosen paradox. *Physics*, 1(3):195–200.

Einstein, A., Podolsky, B. et Rosen, N. (1935). Can quantum-mechanical description of physical reality be considered complete? *Physical Review*, 47: 777–780.

Greenberger, D., Horne, M. et Zeilinger, A. (1989). Going beyond Bell's theorem. *In* Kafatos, M., éditeur : *Bell's Theorem, Quantum Theory and Conceptions of the Universe*, pages 69–72. Kluwer Academic Publishers.

Mermin, N. (1990). *Boojums all the way through*. Cambridge University Press.