Sujet Projet de fin d'études (code 17_1)

Un package R pour l'analyse et la visualisation

des associations et des corrélations et étude de cas

Sujet en entreprise localisée à Jemmel (Monastir)

Contact : Mariem Gzara, mariem.gzara@gmail.com

Année universitaire : 2016/2017

Mots clés : analytique des données, extraction des connaissances, sciences des données, logiciel R, larges bases de données

Développement : Logiciel R (R est aussi un langage orienté objet), C/C++, Java

Des larges volumes de données sont disponibles et accessibles sous diverses formes structurées et non structurées. Cette abondance des données cache une connaissance aujourd'hui primordiale pour la prise de décision stratégique et la commande des systèmes.

L'analyse des associations et des corrélations entre les données rend possible la compréhension du système à travers les données qu'il génère.

L'objectif de ce projet est de :

I. créer un R package qui :
   1) Réunit les algorithmes les plus connus de détections des associations et des corrélations entre les données.
   2) Evalue les connaissances extraites par des métriques de validation.
   3) Visualise les associations et les corrélations à travers des graphiques clairs et compréhensibles par les experts.

II. Utiliser le package crée dans une étude de cas réelle.

Sujet Projet de fin d'études (code 17_2)

Un package R pour la classification supervisée des données

Et étude de cas réelle

Sujet en entreprise localisée à Jemmel (Monastir)

Contact : Mariem Gzara, mariem.gzara@gmail.com

Année universitaire : 2016/2017

Mots clés : analytique des données, extraction des connaissances, logiciel R, classification supervisée.

Développement : Logiciel R (R est aussi un langage orienté objet), C/C++, Java

Des larges volumes de données sont disponibles et accessibles sous diverses formes structurées et non structurées. Cette abondance des données cache une connaissance aujourd'hui primordiale pour la prise de décision stratégique et la commande des systèmes.

Des modèles sont extraits à partir des données pour classifier les nouveaux objets pour lesquels les étiquettes de classes sont non connues. Plusieurs méthodes de classification supervisée existent dont les arbres de décisions, les réseaux de neurones, les algorithmes génétiques, les bases de règles de classification.

L'objectif de ce projet est de :

I. Créer un R package qui :
   1) Réunit les algorithmes et les approches les plus connus de classification supervisée.
   2) Evalue les connaissances extraites, ici les modèles de classification, par des métriques de validation.
   3) Visualise les modèles de classifications (arbres de décisions, bases de règles) à travers des graphiques clairs et compréhensibles par les experts.
II. Utiliser le package crée dans une étude de cas réelle.

Sujet Projet de fin d'études (code 17_3)

Un package R pour la préparation et l'exploration des données

Et étude de cas réelle

Sujet en entreprise localisée à Jemmel (Monastir)

Contact : Mariem Gzara, mariem.gzara@gmail.com

Année universitaire : 2016/2017

Mots clés : analytique des données, extraction des connaissances, logiciel R, préparation et exploration des données

Développement : Logiciel R (R est aussi un langage orienté objet), C/C++, java

Des larges volumes de données sont disponibles et accessibles sous diverses formes structurées et non structurées. Cette abondance des données cache une connaissance aujourd'hui primordiale pour la prise de décision stratégique et la commande des systèmes. Le processus de découverte de la connaissance est un processus itératif qui reçoit les données en entrées et délivre de la connaissance.

La préparation des données est la première étape de ce processus qui nécessite un effort qui peut dépasser les 90% des tâches générées dans un projet datamining. La préparation des données couvre la collecte, la fusion, le nettoyage, la transformation et la réduction. Chaque tâche requière des algorithmes et des fonctions spécifiques.

L'objectif de ce projet est de :

I. Créer un package R pour la préparation des données (importation, exportation, nettoyage, réduction, transformation, etc.
II. Exploiter le package et montrer son efficacité sur des bases de données réelles.

Sujet Projet de fin d'études (code 17_4)

Un package R pour les métriques de similarité et de distance

Et étude de cas réelle

Sujet en entreprise localisée à Jemmel (Monastir)

Contact : Mariem Gzara, mariem.gzara@gmail.com

Année universitaire : 2016/2017

Mots clés : analytique des données, extraction des connaissances, logiciel R,
Développement : Logiciel R (R est aussi un langage orienté objet), C/C++, java

Des larges volumes de données sont disponibles et accessibles sous diverses formes structurées et non structurées. Cette abondance des données cache une connaissance aujourd'hui primordiale pour la prise de décision stratégique et la commande des systèmes. Le processus de découverte de la connaissance est un processus itératif qui reçoit les données en entrées et délivre de la connaissance. La découverte de la connaissance se base sur l'étude des associations, des corrélations et des similarités entre les données.

Les données similaires constituent des catégories, des groupes, des modèles, etc. Une question centrale se pose alors. Comment évaluer la similarité entre les données ? Ces données sont de différents types (continues, discrètes, binaires, catégoriques, etc) et aussi elles sont structurées et non structurées (graphes, séquences, arbres, vecteurs, matrices, images, documents xml, etc). Souvent des métriques de distances sont proposées pour mesurer la dissimilarité.

L'objectif de ce projet est de :

I. Créer un package R qui intègre un large nombre de métriques de calcul de similarité et de dissimilarité
II. Utiliser le package pour évaluer la similarité entre divers types d'objets (images, documents, etc).

Sujet Projet de fin d'études (code 17_5)

Un package R pour la fouille des larges graphes et étude de cas réelle

Sujet en entreprise localisée à Jemmel (Monastir)

Contact : Mariem Gzara, mariem.gzara@gmail.com

Année universitaire : 2016/2017

Mots clés : analytique des données, extraction des connaissances, logiciel R,
Développement : Logiciel R (R est aussi un langage orienté objet), C/C++, java

Des larges volumes de données sont disponibles et accessibles sous diverses formes structurées et non structurées. Cette abondance des données cache une connaissance aujourd'hui primordiale pour la prise de décision stratégique et la commande des systèmes. Or, l'extraction des connaissances est basée sur la modélisation de ces volumes des données. Les graphes sont des modèles proches des situations réelles où des entités sont reliées entre elles par diverses liaisons telles que l'appartenance d'un individu à un organisme ou les relations d'amitiés ou de centres d'intérêts commun.

Ce projet a pour objectif d'étudier les graphes comme outil de modélisation des données massives à travers quelques exemples, d'explorer les codages efficaces de ces graphes et d'implémenter efficacement des algorithmes d'extraction des connaissances dans des larges graphes. La validation du travail mené se fait à travers une étude de cas réelle.

Le travail demandé consiste à créer un package R qui :

- Intègre des types de codage multiples des graphes et des hypergraphes.
- Englobe des algorithmes d'extraction des connaissances dans les graphes.
- Valide les algorithmes implémentés par des métriques de validation et par visualisation.
- Exploite le package dans une étude de cas réelle.

Sujet Projet de fin d'études (code 17_6)

Un package R pour la sélection des attributs et étude de cas réelle

Sujet en entreprise localisée à Jemmel (Monastir)

Contact : Mariem Gzara, mariem.gzara@gmail.com

Année universitaire : 2016/2017

Mots clés : analytique des données, extraction des connaissances, logiciel R,
Développement : Logiciel R (R est aussi un langage orienté objet), C/C++, java

Des larges volumes de données sont disponibles et accessibles sous diverses formes structurées et non structurées. Cette abondance des données cache une connaissance aujourd'hui primordiale pour la prise de décision stratégique et la commande des systèmes. Or, cette connaissance se trouve noyée par deux difficultés à savoir la numérosité et la dimensionnalité. Le premier obstacle est relatif aux nombres d'observations (Big data) et le deuxième est relatif aux nombres très élevé des variables. Dans une base de données, un attribut est une variable. Dans les situations réelles, ce nombre d'attributs atteint facilement plusieurs centaines d'attributs. Ces attributs ne sont pas tous pertinents et leur considération simultané peut biaiser le comportement des algorithmes comme elle peut rendre impossible l'exécution d'autres. La réduction de la dimension par la sélection des attributs pertinents est à la fois un prétraitement et une fonctionnalité d'extraction des connaissances à part entière.

L'objectif de ce projet est de créer un package R qui :

- Englobe des algorithmes de sélection des attributs implémentés efficacement.
- Valide les algorithmes de sélection par des métriques numériques et par visualisation.
- Exploite le package crée dans une étude de cas réelle.

Sujet Projet de fin d'études (code 17_7)

Un package R pour la classification automatique

Sujet en entreprise localisée à Jemmel (Monastir)

Contact : Mariem Gzara, mariem.gzara@gmail.com

Année universitaire : 2016/2017

Mots clés : analytique des données, extraction des connaissances, logiciel R,
Développement : Logiciel R (R est aussi un langage orienté objet), C/C++, java

Des larges volumes de données sont disponibles et accessibles sous diverses formes structurées et non structurées. Cette abondance des données cache une connaissance aujourd'hui primordiale pour la prise de décision stratégique et la commande des systèmes. Le clustering est fonctionnalité centrale en fouilles de données parce qu'elle découvre des catégories dans les données qui permettent une meilleure description des données. Une large gamme d'algorithmes de classification automatique est proposée. Il n'y a pas un algorithme qui devance les autres. Il est essentiel de pouvoir discriminer entre plusieurs algorithmes de clustering soit en effectuant un choix au départ, soit en comparant les résultats, soit en combinant les résultats.

L'objectif de ce projet est de :

- créer un package R qui :
    - Intègre plusieurs algorithmes de clustering.
    - Intègre des algorithmes d'agrégation de clusters.
    - Génère un benchmark de bases données artificielles
- Valider le package créer sur une étude de cas réelle

Sujet Projet de fin d'études (code 17_8)

Un package R pour la recommandation des objets et étude de cas réelle

Sujet en entreprise localisée à Jemmel (Monastir)

Contact : Mariem Gzara, mariem.gzara@gmail.com

Année universitaire : 2016/2017

Mots clés : analytique des données, extraction des connaissances, logiciel R, système de recommandation

Développement : Logiciel R (R est aussi un langage orienté objet), C/C++, java

Les systèmes de recommandation sont intégrés dans plusieurs services comme le commerce électronique pour la recommandation des articles aux utilisateurs potentiels. Ces recommandations sont extraites à partir des volumes de données décrivant les articles, les utilisateurs et leur comportement. Les approches de recommandations adoptent trois stratégies à savoir la recommandation personnalisée, recommandation Objet (Content-Based filtering CB), recommandation Sociale (Collaborative Filtering CF–Context Aware) et la recommandation Hybride qui combine ses approches. Un nombre élevé d'algorithmes ont été proposés dans ces systèmes de recommandation.

L'objectif de ce projet est de créer un package R qui :

- Implémente le processus de recommandation
- Intégrer différents algorithmes de recommandation.

En deuxième lieu, l'étudiant doit valider le package dans une étude de cas réelle.

Sujet Projet de fin d'études (code 17_9)

Un package R pour la collecte des données

Sujet en entreprise localisée à Jemmel (Monastir)

Contact : Mariem Gzara, mariem.gzara@gmail.com

Année universitaire : 2016/2017

Mots clés : analytique des données, extraction des connaissances, logiciel R, système de recommandation

La science des données est la science du 21ème siècle. C'est la science qui va bouleverser notre vision des systèmes, de leurs évolutions et de leurs commandes. Cette science touche toutes les fonctionnalités de traitement automatique des données. Les données sont abondantes, distribuées, structurées, non structurées, hétérogènes, etc. La première étape clé est comment collecter ces données à partir des bases de données distribuées et hétérogènes, à partir du web, à partir des réseaux sociaux, des documents, des émails, etc.

L'objectif de ce projet est de :

- créer un package R pour la collecte puis le stockage des données.
- valider le package sur des sources de données réelles.