Institut Francophone International

RAPPORT

FOUILLE DE DONNÉES

“Bilan des logiciels de fouille de données”

Étudiante : DAO Thuy Hong
Promotion : 20

Hanoi, Décembre 2016

Sommaire

Les logiciels de fouille de données sont des programmes spécialisés dans l'analyse et l'extraction des connaissances à partir des données informatisées. Ce sont des logiciels qui aident l'analyste en exploration de données à trouver des motifs remarquables et intéressants. Il peut s'agir de logiciels commerciaux ou de logiciels libres.

1. Logiciels libres :

1.1 KNIME :

KNIME (prononcer NAÏM), acronyme de Konstanz Information Miner1, est un logiciel libre édité par un laboratoire de l'université de Constance dénommé Nycomed Chair for Bioinformatics and Information Mining2,3.

Figure 1 : Logo de KNIME

KNIME permet aux utilisateurs de créer visuellement des flux de données (ou pipelines), d'exécuter sélectivement certaines ou toutes les étapes d'analyse, puis d'inspecter les résultats, les modèles et les vues interactives. KNIME est écrit en Java et basé sur Eclipse et utilise son mécanisme d'extension pour ajouter des plugins fournissant des fonctionnalités supplémentaires. La version de base inclut déjà des centaines de modules pour l'intégration des données (E/S de fichiers, nœuds de bases de données prenant en charge tous les systèmes de gestion de base de données courants), la transformation des données (filtre, convertisseur, combinateur) ainsi que les méthodes couramment utilisées pour l'analyse et la visualisation des données. Avec l'extension gratuite Report Designer, les flux de travail KNIME peuvent être utilisés comme ensembles de données pour créer des modèles de rapports qui peuvent être exportés vers des formats de document comme doc, ppt, xls, pdf et autres. Les autres capacités de KNIME sont:

- L'architecture de base de KNIME permet de traiter des volumes de données volumineux qui ne sont limités que par l'espace disque disponible (la plupart des autres outils d'analyse de données open source fonctionnent dans la mémoire principale et sont donc limités à

la RAM disponible). Par exemple. KNIME permet d'analyser 300 millions d'adresses clients, 20 millions d'images cellulaires et 10 millions de structures moléculaires.

- Des plugins supplémentaires permettent d'intégrer des méthodes pour l'exploration de texte, l'exploration d'images et l'analyse de séries chronologiques.
- KNIME intègre divers autres projets libres : les algorithmes d'apprentissage de la machine de Weka, le projet R, aussi bien que LIBSVM, JFreeChart, ImageJ, et Chemistry Development Kit.

KNIME est implémenté en Java mais permet également aux wrappers d'appeler d'autres codes en plus de fournir des nœuds permettant d'exécuter Java, Python, Perl et d'autres fragments de code.

1.2 Orange :

Orange est un logiciel libre créé à l'université de Ljubljana en Slovénie. Ce logiciel est doté d'une interface homme-machine conviviale. Il est développé en C++ et en Python. Chaque algorithme se présente sous la forme de widgets pouvant avoir une entrée et une sortie ; ils sont agencés dans une fenêtre.

Figure 2 : Logo d'orange.

Caractéristiques :

- **Canvas :** font-end graphique pour l'analyse de données.
- **Widgets :**
  - ✓ **Data :** Widgets pour l'entrée de données, le filtrage des données, l'échantillonnage, l'imputation, la manipulation des fonctions et la sélection des fonctions.
  - ✓ **Visualize :** Widgets pour une visualisation commune (encadré, histogrammes, diagramme de dispersion) et visualisation multivariée (affichage mosaïque, diagramme tamis).

- ✓ **Classify :** Un ensemble d'algorithmes d'apprentissage machine supervisés pour la classification.
- ✓ **Regression** : Un ensemble d'algorithmes d'apprentissage machine supervisés pour la regression.
- ✓ **Evaluate :** Validation croisée, procédures basées sur l'échantillonnage, estimation de la fiabilité et notation des méthodes de prevision.
- ✓ **Unsupervised :** Algorithmes d'apprentissage non supervisés pour le clustering (k-means, clustering hiérarchique) et techniques de projection de données (échelle multidimensionnelle, analyse des composantes principales, analyse de correspondance).

1.3 Weka :

Weka (Waikato Environment for Knowledge Analysis) est une suite populaire de logiciels d'apprentissage machine écrit en Java, développé à l'Université de Waikato, en Nouvelle-Zélande. Il s'agit d'un logiciel libre sous licence GNU General Public License.

Figure 3 : Logo de Weka

Weka est une collection d'algorithmes d'apprentissage automatique pour les tâches d'exploration de données. Les algorithmes peuvent être appliqués directement à un ensemble de données ou appelés à partir de votre propre code Java. Weka contient des outils pour le prétraitement des données, la classification, la régression, le regroupement, les règles d'association et la visualisation. Il est également bien adapté pour le développement de nouveaux systèmes d'apprentissage automatique.

Weka fournit l'accès aux bases de données de SQL using Java Database Connectivity et peut traiter le résultat retourné par une requête de base de données. Il n'est pas capable de l'exploration de données multi-relationnelle,

mais il existe un logiciel séparé pour convertir une collection de tables de base de données liées dans une seule table qui convient au traitement à l'aide de Weka. Un autre domaine important qui n'est actuellement pas couvert par les algorithmes inclus dans la distribution de Weka est la modélisation de séquences.

1.4 Tanagra :

Tanagra est une suite gratuite de logiciels d'apprentissage automatique conçus par Ricco Rakotomalala à l'Université Lumière Lyon 2, France. Tanagra prend en charge plusieurs tâches d'exploration de données standard telles que : visualisation, statistiques descriptives, sélection d'instance, sélection d'entités, construction de caractéristiques, régression, analyse factorielle, regroupement, classification et apprentissage des règles d'association.

Tanagra fonctionne comme les outils de data mining actuels. L'utilisateur peut concevoir visuellement un processus d'exploration de données dans un diagramme. Chaque noeud est une technique statistique ou d'apprentissage machine, la connexion entre deux noeuds représente le transfert de données. Mais contrairement à la majorité des outils basés sur le paradigme du workflow, Tanagra est très simplifié. Les traitements sont représentés dans un diagramme en arbre. Les résultats sont affichés en format HTML. Il est donc facile d'exporter les sorties pour visualiser les résultats dans un navigateur. Il est également possible de copier les tableaux de résultats dans une feuille de calcul.

2. Logiciels commerciaux :

2.1 SAS Enterprise Miner :

SAS (Statistical Analysis System) [1] est une suite logicielle développée par SAS Institute pour l'analyse avancée, les analyses multivariées, l'intelligence d'affaires, la gestion des données et l'analyse prédictive.

SAS a été développé à l'Université d'Etat de Caroline du Nord de 1966 à 1976, quand l'Institut SAS a été incorporé. SAS a été développé dans les années 1980 et 1990 avec l'ajout de nouvelles procédures statistiques, des composants supplémentaires et l'introduction de JMP. Une interface point-and-click a été ajoutée en version 9 en 2004. Un produit d'analyse des médias sociaux a été ajouté en 2010.

Figure 4 : Logo de SAS.

SAS est une suite logicielle qui peut exploiter, modifier, gérer et extraire des données à partir d'une variété de sources et d'effectuer des analyses statistiques. SAS fournit une interface graphique point-and-click pour les utilisateurs non techniques et des options plus avancées via le langage SAS. Pour utiliser Statistical Analysis System, les données doivent être dans un format de tableur ou un format SAS. Les programmes SAS ont une étape DATA, qui récupère et manipule des données, créant habituellement un ensemble de données SAS, et une étape PROC, qui analyse les données.

Chaque étape consiste en une série d'énoncés. L'étape DATA comporte des instructions exécutables qui permettent au logiciel de prendre une action et des instructions déclaratives qui fournissent des instructions pour lire un ensemble de données ou altérer l'apparence des données. L'étape DATA comporte deux phases, la compilation et l'exécution. Dans la phase de compilation, les déclarations sont traitées et les erreurs de syntaxe sont identifiées. Ensuite, la phase d'exécution traite chaque instruction exécutable séquentiellement. Les ensembles de données sont organisés en tables avec des lignes appelées «observations» et des colonnes appelées «variables». De plus, chaque élément de données possède un descripteur et une valeur.

L'étape PROC consiste en des instructions PROC appelant des procédures nommées. Les procédures permettent d'effectuer des analyses et des rapports sur des ensembles de données afin de produire des statistiques, des analyses et des graphiques. Il existe plus de 300 procédures et chacune contient un ensemble important de programmes et de travaux statistiques. Les instructions PROC peuvent également afficher des résultats, trier des données ou effectuer d'autres opérations. Les macros SAS sont des morceaux de code ou des variables qui sont codées une fois et référencées pour exécuter des tâches répétitives.

2.2 Statistica Data Miner :

Statistica est un progiciel d'analyse avancée développé à l'origine par StatSoft qui a été acquis par Dell en mars 2014. Statistica fournit des analyses de données, de la gestion des données, des statistiques, de l'exploration de données, de l'apprentissage automatique, de l'analyse de texte et des procédures de visualisation de données. Les catégories de produits Statistica incluent l'entreprise (pour une utilisation sur un site ou une organisation), sur le Web (pour un serveur et un navigateur Web), sur un réseau concurrent et sur un bureau mono-utilisateur.

Figure 5 : Logo de Statistica.

2.3 Braincube :

Braincube, de la société IP Leanware, est une solution cloud leader mondial sur le marché émergent de l'Operational Intelligence qui intègre des algorithmes en grille pour la mesure des impacts entre variables et la recherche de solution de réglages optimaux. C'est la première solution bigdata utilisée dans les usines de production de masse.

Figure 6 : Logo de Braincube

Figure 7 : Caractéristiques de Braincube

3. Logiciels specializes :

En fouille de données spatiales, les logiciels sont aptes à analyser, requêter et tenir compte des spécificités des données spatiales. Il existe quelques logiciels comme : Geominer, Descartes, Fuzzy Spatial OQL for Fuzzy KDD, GWiM GeoKD 6, SPIN! 7, S+SpatialStats, Modules spécialisés dans R, Spatial Statistics Toolbox for Matlab/Fortran, NEM *.

3.1 S+SpatialStats :

S + SPATIALSTATS est le premier paquet complet et orienté objet pour l'analyse des données spatiales. Fournissant un nouvel ensemble d'outils d'analyse, S + SPATIALSTATS a été créé spécifiquement pour l'exploration et la modélisation de données spatialement corrélées.

Profitant pleinement des méthodes orientées objet et du langage de modélisation de S-PLUS, S + SpatialStats nous permet d'analyser les données avec une structure spatiale complète et correcte. S + SpatialStats peut être utilisé pour analyser des données provenant de domaines techniques tels que l'environnement, l'ingénierie minière et pétrolière, les ressources naturelles, la géographie, l'épidémiologie, la démographie et d'autres où les données sont échantillonnées spatialement.

3.2 Spatial Statistics Toolbox for Matlab/Fortran :

Spatial Statistics Toolbox for Matlab/Fortran comprend le code pour les autorégressions spatiales simultanées (SAR), les autorégressions spatiales conditionnelles (CAR) et les modèles régressifs à régression spatiale autorégressive (MRSA). En outre, il contient un code pour créer des matrices de poids spatials clairsemés et trouver les log-déterminants (nécessaires pour le maximum de vraisemblance). Par conséquent, la Matlab Spatial Statistics Toolbox comprend les estimateurs les plus courants utilisés dans l'économétrie spatiale. Ces produits utilisent des matrices dispersées et d'autres techniques de calcul pour accélérer considérablement les calculs et augmenter la taille des ensembles de données potentielles analysés.

Référence :

[1] https://fr.wikipedia.org/wiki/Logiciels_de_fouille_de_donn%C3%A9es

[2] https://www.knime.org

[3] http://www.ailab.si/orange

[4] http://www.springer.com/us/book/9780387982267

[5] http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html

[6] http://www.spatial-statistics.com/software_index.htm

[7] http://braincube.com/

[8] http://www.sas.com/en_us/software/analytics/enterprise-miner.html

[9] http://www.cs.waikato.ac.nz/ml/weka/