Les logiciels gratuits pour l'enseignement du Data Mining

Ricco RAKOTOMALALA
Laboratoire ERIC
Université Lumière Lyon 2
http://eric.univ-lyon2.fr/~ricco

Pour ou contre les logiciels gratuits dans l'enseignement du Data Mining ?

1. S'attacher au fond (*méthode - savoir*) et non pas à la forme (*savoir-faire*)

Qu'importe le logiciel et son mode opératoire ?

2. Former des étudiants qui iront sur le marché du travail

Il existe des « standards » selon les domaines, les logiciels commerciaux les respectent, il faut

Familiariser les étudiants avec ces standards

3. L'apprentissage d'un logiciel ne doit pas requérir des compétences supplémentaires spécifiques (hors domaine)

Accès à des formats de fichiers reconnus, langage de script spécifique

Cahier des charges pour les logiciels utilisables dans l'enseignement du Data Mining

1. Gratuité sans restrictions – Au moins dans le cadre des enseignements
2. Installation simplifiée – Pas de serveurs lourds à installer
3. Gestion simplifiée des données -- Format texte / tableur
4. Fonctionnement par diagramme de traitements – La fameuse notion de « filière »
5. Évaluation des méthodes (supervisées), outils de scoring et comparaisons
6. Résultats lisibles, possibilité de les reprendre dans un traitement de texte

Pouvoir définir des traitements, les comparer, les évaluer sur des jeux de données sans avoir à passer par un apprentissage compliqué d'un logiciel spécifique

Les critères (logiciels commerciaux) éludés

1. Interfaçage avec les bases de données
2. Traitement à la volée sur de très grandes bases de données (pas de données en mémoire)
3. Déploiement des modèles construits et mise en production
4. Reporting évolué et dynamique (màj à la volée dans les documents édités)
5. Exploration graphique évoluée et interactive (isoler graphiquement des sous-population)
6. *… le prix … ?*

Quels logiciels gratuits ?

Référence : le site KDNUGGETS (Software – Suites – Free) → http://www.kdnuggets.com/software/suites.html

- ADaM, Algorithm Development and Mining version 4.0 toolkit
- AlphaMiner, open source data mining platform that offers various data mining model building and data cleansing functionality.
- Databionic ESOM Tools, a suite of programs for clustering, visualization, and classification with Emergent Self-Organizing Maps (ESOM).
- Gnome Data Mining Tools, including apriori, decision trees, and Bayes classifiers.
- IBM Intelligent Miner. University scholars can now receive free copies of DB2 UDB and Intelligent Miner for educational or research purposes.
- MiningMart, a graphical tool for data preprocessing and mining on relational databases; supports development, documentation, re-use and exchange of complete KDD processes. Free for non-commercial purposes.
- MLC++, a machine learning library in C++.
  See also Kansas State U. port of MLC++: Binary (tar.gz), and Linux source
- Machine Learning in Java (MLJ), an open-source suite of Java tools for research in machine learning.
- Orange, C++ components for data mining,includes preprocessing, modelling and data exploration techniques.
- StarProbe, Web-based multi-user server available for academic institutions.
- Superinduction, based on SPSS Clementine and other methods.
- TANAGRA, offers a GUI interface and methods for data access, statistics, feature selection, classification, clustering, visualization, association and more.
- Weka, collection of machine learning algorithms for solving real-world data mining problems. It is written in Java and runs on almost any platform.
- YALE (Yet Another Learning Environment), flexible and modular environment for machine learning, including classification, regression, and clustering; XML config; and more ...

Quels critères d'évaluation ?

- Origine, architecture : le mode opératoire, la philosophie
- Interface : chaîne de traitements, graphe ou arbre
- Accès aux données : fichier texte, autres …
- Enchaîner les traitements : exploiter les variables intermédiaires dans d'autres méthodes
- Bibliothèque de méthodes : évaluer l'étendue des méthodes disponibles
- Évaluation et comparaisons : dans le sens comparer les approches (méthodes, construction de var.) pour le supervisé
- Expérimentations : monter des expérimentations (paramétrées) à grande échelle pour les publications
- Performances : capacités de traitement (individus x variables), temps de calcul
- Exploration graphique : visualisation des données, possibilité d'interactivité

WEKA – ORANGE – TANAGRA

Des outils riches mais des objectifs, des cultures et des approches très différentes

WEKA

Un outil privilégié pour les expérimentations
et la comparaison de performances en supervisé

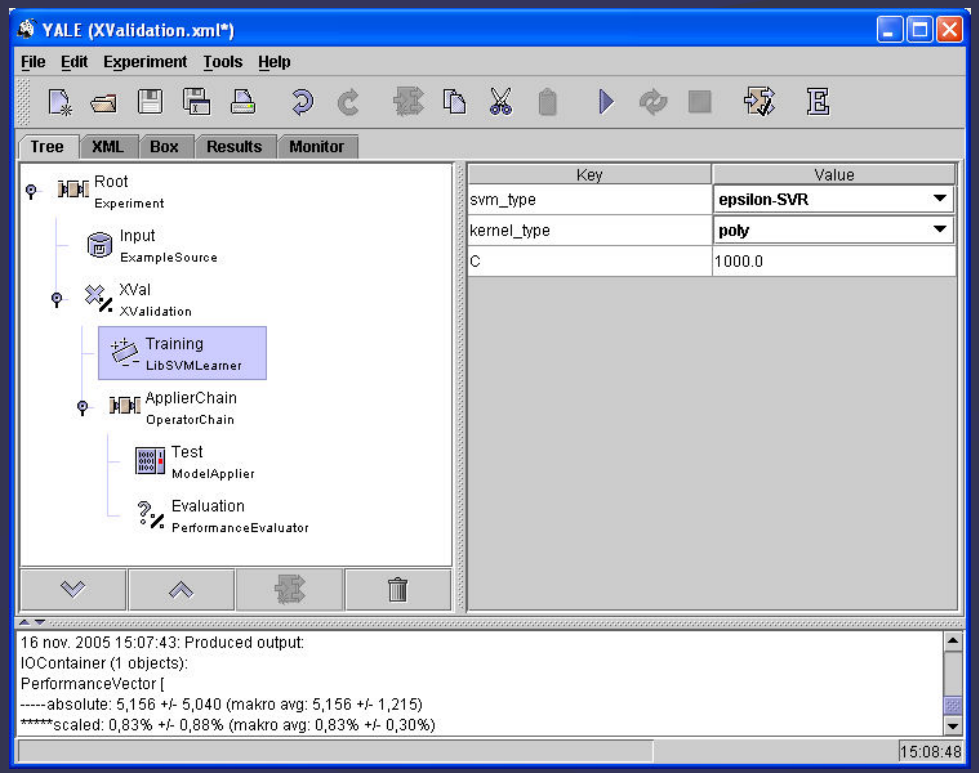
ORANGE

Un outil très marqué « machine learning », souple
avec des efforts pour la convivialité et la simplicité

TANAGRA

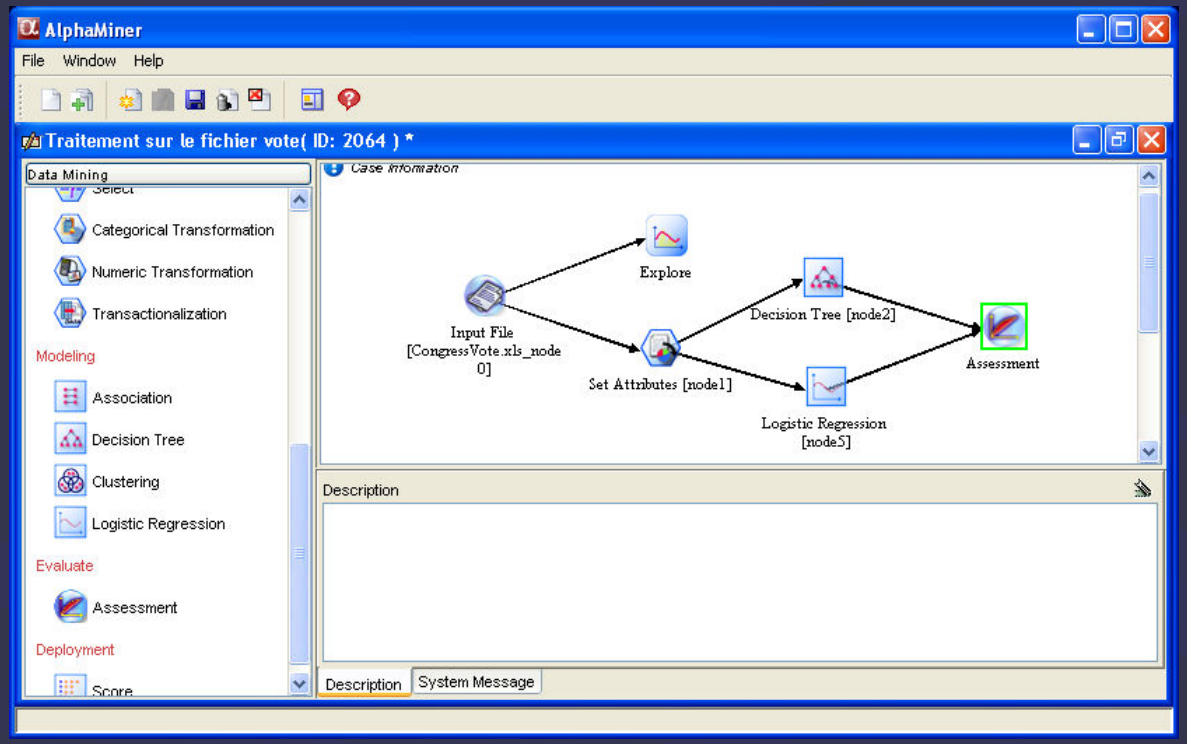
Un outil pluri-culturel tourné vers les études
de statistique exploratoire

YALE



- Origine, architecture : Université de Dortmund, Utilisation des bibliothèques WEKA, culture DATA MINING
- Interface : chaîne de traitements linéaire, une branche sert à définir un sous-traitement (autre : WIZARD ou XML), résultats très frustres
- Accès aux données : fichier WEKA et XML (sur couche WEKA avec définition plus précise des attributs)
- Enchaîner les traitements : limitées au « data préparation », utilisables par la suite
- Bibliothèque de méthodes : idem WEKA
- Évaluation et comparaisons : oui pour l'évaluation (possibilités très élaborées), difficile pour la comparaison
- Expérimentations : ???, il semble que non
- Performances : limitations de JAVA
- Exploration graphique : limitée, calquée sur WEKA

ALPHA MINER



- Origine, architecture : Université de Hong-Kong, Utilisation des bibliothèques WEKA et XELOPES, culture DATA MINING
- Interface : chaîne de traitements sous forme de graphes, interface simplifiée, sorties en PMML
- Accès aux données : fichier texte (séparateur « , ») et XLS, autres texte
- Enchaîner les traitements : limitées au « data préparation », utilisables par la suite
- Bibliothèque de méthodes : volontairement limitée (simplifiée) mais extensible (cf. WEKA par ex.)
- Évaluation et comparaisons : oui, possibilité de comparer les taux de classement et les courbes LIFT
- Expérimentations : ???, il semble que non
- Performances : limitations de JAVA
- Exploration graphique : limitée, calquée sur WEKA

CONCLUSION – BILAN

Oui pour les aspects méthodologiques

- les interfaces sont dans les « standards » actuels (cf. SPAD, SAS-EM, SPSS-CLEM, S-PLUS INSIGHT FULL MINER, STATISTICA DATA MINER…)
- il s'agit de montrer/comprendre le fonctionnement des méthodes
- lire et interpréter les résultats
- comparer les méthodes sur les jeux de données
- richesse des méthodes disponibles et la connexion avec la recherche

Non pour les aspects opérationnels

- l'interfaçage avec les bases de données et l'exploration « on-line » de gros volumes
- l'exploration interactive et graphique, le retour à la base de données
- le « reporting » de qualité et dynamique
- la mise en production pour exploiter les résultats

En fait tout ce qui demande beaucoup d'ingénierie et peu rentable en « publications »