INTRODUCTION AU DATA MINING AVEC R : VERS LA RECONQUÊTE DU «KNOWLEDGE DISCOVERY IN DATABASES» PAR LES STATISTICIENS

Dr. Diego Kuonen, kuonen@statoo.com, Statoo Consulting,
PO Box 107, CH-1015 Lausanne

R est un langage et un environnement pour les calculs statistiques et leurs représentations graphiques. R est similaire au système S qui est la plate-forme du logiciel commercial S-PLUS.

Le but initial de cet article est de proposer un point de départ pour les novices intéressés par R. Après une courte introduction, l'article se propose d'illustrer les interfaces existantes entre R et les bases de données relationnelles, ces interfaces étant un premier pas des logiciels statistiques modernes, comme R, vers la reconquête par les statisticiens du domaine du «Knowledge Discovery in Databases» (KDD).

Cet article est essentiellement basé sur trois articles ([1], [2] et [3]) écrits pour le «Flash Informatique» de l'EPFL.

Qu'est-ce que R?

R est un système d'analyse statistique créé par Ross Ihaka et Robert Gentleman [4] du Département de Statistique de l'Université de Auckland. Depuis le début, un grand nombre de personnes ont également contribué à R en programmant. R est à la fois un langage et un logiciel. Parmi ses caractéristiques les plus remarquables, citons,

- un système performant de stockage et de manipulation de données;
- la possibilité d'effectuer des calculs matriciels et autres opérations complexes;
- une large collection intégrée et cohérente d'outils d'analyse statistique;
- un large éventail d'outils graphiques particulièrement flexibles;
- un langage de programmation simple et efficace qui include nombreuses facilités.

R est un langage qualifié de dialecte du langage S créé par Lucent Technologies (AT&T Bell Labs). S est disponible sous la forme du logiciel S-PLUS commercialisé par la compagnie américaine Insightful Corporation. Il existe quelques différences entre la conception de R et celle de S. Dans cet article, je ne m'attarde pas sur ces différences reconnues importantes. Le lecteur intéressé peut se référer à «R FAQ» [5] dont une copie se trouve avec le logiciel.

Redistribué librement sous les termes de la «GNU Public Licence» de la «Free Software Foundation», le développement et la distribution de R sont assurés par plusieurs statisticiens rassemblés dans le «R Development Core Team» qui existe depuis mi-1997 et qui modifie l'archive CVS de la source du code R. Un élément essentiel du développement est le «Comprehensive R Archive Network» [6] dont un exemplaire se trouve à l'ETHZ [6], ainsi que sur le site officiel de R [7]. De plus, notons aussi que R fait partie officiellement du projet GNU («GNU S»).

R est disponible sous plusieurs formes: code écrit en C (et certaines routines en Fortran) prêt à être compilé, surtout pour les machines Unix/GNU Linux, ou exécutables prêts à l'emploi pour Windows et pour MacOS(X).

Le noyau de R est un langage de programmation interprété qui permet les embranchements et les boucles aussi bien que la programmation modulaire utilisant des fonctions. La plupart des fonctions visibles par l'utilisateur dans R sont écrites en R. L'interface avec des procédures écrites en langage C, C++ ou Fortran est possible pour l'utilisateur qui gagne ainsi en efficacité. La distribution de R contient une fonctionnalité pour un grand nombre de procédures statistiques. Il existe également beaucoup de fonctions qui fournissent un environnement graphique très flexible pour créer divers types de présentations des données. Les modules supplémentaires («add-on packages») sont nombreux et sont disponibles pour des buts spécifiques (pour une liste, voir [5] et [6]).

R est un langage qui comporte de nombreuses fonctions pour les analyses statistiques et les graphiques; ces derniers se situent dans des fenêtres spécifiques et peuvent être exportés sous divers formats. Les résultats des calculs statistiques sont affichés à l'écran et certains résultats partiels peuvent être sauvés ailleurs, exportés dans un fichier ou utilisés pour des analyses ultérieures. Le langage R permet, par exemple, de programmer des boucles qui vont ana-

lyser successivement divers jeux de données. Il est aussi possible de combiner dans le même programme différentes fonctions statistiques pour réaliser des analyses plus complexes. Les utilisateurs de R peuvent bénéficier des nombreuses routines écrites pour S-PLUS, la plupart d'entre elles étant directement utilisables par R.

De prime abord, R peut sembler trop complexe pour une utilisation par un non-spécialiste. Ce n'est pas forcément le cas. En fait, R privilégie la flexibilité. Alors qu'un logiciel classique (SAS, SPSS, Statistica, ...) affiche directement tous les résultats (ou presque) d'une analyse, dans R, ils sont stockés dans un objet, si bien qu'une analyse peut être faite sans qu'aucun résultat n'apparaisse à l'écran. L'utilisateur peut en être perturbé, mais ceci se révèle extrêmement utile. En effet, l'utilisateur peut alors extraire uniquement une partie des résultats qui l'intéressent.

Une fois R installé sur votre ordinateur, il suffit de lancer l'exécutable correspondant pour accéder au programme. L'aide en ligne de R donne de très bonnes informations sur l'utilisation des fonctions. Notons aussi que R possède sa propre documentation disponible sur la toile [6] sous différents formats ou en forme brute. La distribution de R s'accompagne également de différents manuels [6]. En complément au matériel écrit spécialement pour R, la documentation de S/S-PLUS peut être utilisée en combinaison avec le «R FAQ» [5]. Aussi, vous trouverez toujours une aide précieuse dans les divers «mailing lists» de R [7] et dans ses archives. Et il y a aussi «R News» [7] destiné à combler le décalage entre les «mailing lists» et les journaux scientifiques.

Visualisation à l'aide de R

Plusieurs personnes utiliseront R principalement pour ses outils graphiques. Ces derniers sont une composante extrêmement souple de l'environnement R. Il est possible de les utiliser pour représenter une grande variété de graphes statistiques mais aussi pour construire entièrement de nouveaux types de graphiques. Les outils graphiques peuvent être utilisés soit en interactif soit en mode queue, bien que dans beaucoup de cas l'utilisation interactive soit plus productive. L'utilisation interactive est aussi plus facile car au lancement de R, une fenêtre graphique s'ouvre pour afficher les graphes interactifs.

Une fois le dispositif enclenché, les commandes de représentations graphiques de R peuvent être utilisées pour produire une grande variété de graphes ainsi que de nouveaux types de graphes.

Une des forces de R est la facilité à produire des graphiques de haute qualité, incluant si nécessaire des formules ou des symboles mathématiques. Bien que le choix des détails graphiques par défaut soit satisfaisant, l'utilisateur détient le contrôle total de la conception de son graphe.

Par exemple, la Figure 1 est un graphe obtenu en utilisant R.

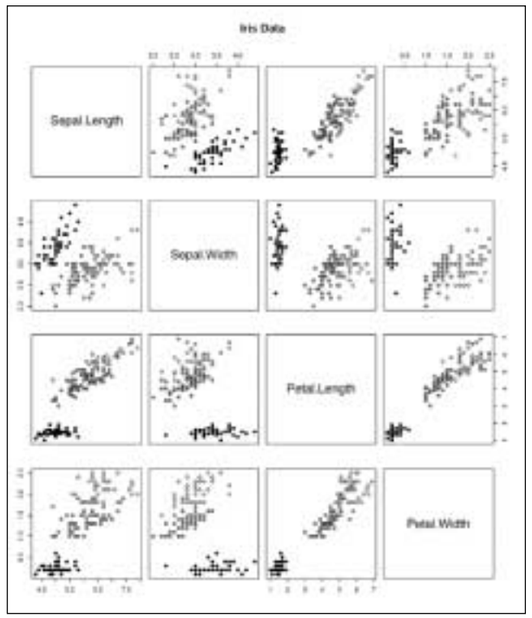


Figure 1. «Iris Data». Représentation graphique de toutes les paires de variables (Figure 1 de [1]).

Qu'est-ce que le data mining?

"We are drowning in information, but starving for knowledge."
(John Naisbett dans son livre "Megatrends")

De manière générale, on peut définir le «Data Mining» (ou «Exploitation des Gisements de Données») comme l'extraction, à partir de gros volumes de données, d'informations ou de connaissances originales, auparavantl inconnues, potentiellement utiles.

Selon SAS Institute, il s'agit du processus de sélection, d'exploration, de modification et de modélisation de grandes bases de données afin de découvrir des relations entre les données jusqu'alors inconnues.

Le data mining correspond donc à l'ensemble des techniques et des méthodes qui, à partir de gros volumes de données, permettent d'obtenir des connaissances exploitables. Son utilité est grande dès qu'une entreprise possède un grand nombre d'informations stockées sous forme de bases de données.

Il existe une distinction précise entre le concept de KDD («Knowledge Discovery in Databases» ou «Découverte de Connaissances dans les Bases de Données») et celui de data mining. En effet, ce dernier n'est que l'une des étapes du processus de découverte de connaissances correspondant à l'extraction des connaissances à partir des données. Pour pouvoir réaliser des études de data mining il faut d'abord disposer d'un «Data Warehouse» («Entrepôt de Données»).

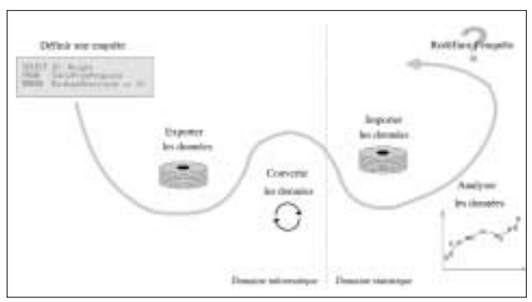
Les applications du data mining sont multiples: la grande distribution, la vente par correspondance, les opérateurs de télécommunication, les banques et les assurances, l'étude des génomes dans la bioinformatique, par exemple, en trouvant des gènes dans des séquences d'ADN, etc. Le domaine principal où le data mining a prouvé son efficacité est la gestion de la relation client, CRM («Customer Relationship Management»). En effet, dans ce cas, le data mining permet d'accroître les ventes par une meilleure connaissance de la clientèle.

Le data mining et ses outils, bien qu'utilisant une stratégie et des techniques statistiques, sont appelés à être utilisés par des non-statisticiens. Regardons pourquoi.

Le domaine de la statistique, spécialement la statistique dite mathématique, a commencé à croître il y a environ 100 ans. A l'époque les jeux de données étaient de petite taille et les calculs s'effectuaient manuellement. L'ère de la statistique mathématique s'est «terminée» le jour où, grâce aux ordinateurs, la statistique dite computationnelle est apparue. Cependant, le stockage électronique des données restait toujours coûteux et les volumes des données toujours limités. Et, comme les statistiques font traditionnellement partie du domaine des mathématiques et non pas de l'informatique, la statistique trouvait son intérêt premier dans la théorie mathématique sous-jacente, et non pas dans les aspects du calcul et du stockage de données à traiter. Ceci explique pourquoi les logiciels statistiques étaient habituellement bons en lecture, respectivement en importation de données depuis un fichier, mais ont malheureusement ignoré pendant trop longtemps le fait que des données reposaient dans des bases de données et l'intérêt qu'il y avait de s'interfacer avec les bases de données. Depuis le début des années 1990, le nombre de très grandes bases de données n'a pas cessé de croître. Les agences gouvernementales, les grands détaillants ou les commerçants en ligne font face maintenant à des énormes bases de données, difficiles à analyser au delà de simples indicateurs sommaires. Ainsi, les informaticiens ont tiré profit de ce manque de connaissances de base de données des statisticiens et ont inventé une discipline appelée KDD. La question reste posée: cette nouvelle discipline est-elle en concurrence avec la statistique?

Puisque KDD a pour but de recueillir des données, de prélever des données, d'autoriser la conception expérimentale et d'analyser des données en utilisant les modèles de découverte, de visualisation, de faire des prévisions et des classifications, elle se situe clairement dans le domaine de la statistique. En résumé, on peut dire que «KDD = Statistics at Scale and Speed». Mais, il reste encore à la statistique de relever les défis du data mining.

Le processus usuel du data mining est représenté dans la Figure 2 et peut se résumer ainsi:



1. définir une requête pour extraire toutes les données souhaitées de la base de données;
2. exporter les données sélectionnées dans un fichier;
3. si nécessaire, les convertir dans un format qui facilite l'importation dans un logiciel statistique tel que R;
4. importer les données dans le logiciel;
5. analyser les données dans le logiciel.

Il faut recommencer le processus tant que l'on n'a pas extrait toutes les données appropriées ou bien si les données fondamentales dans la base de données changent, ce qui nécessite la mise à jour des résultats.

La possibilité de manipuler des données se trouvant dans des bases de données est très importante mais est absente dans beaucoup de logiciels statistiques. Travailler directement sur des bases de données avec des outils interactifs semble être beaucoup plus difficile que travailler sur des données structurées spécialement pour des applications statistiques. Mais c'est la seule possibilité qui s'offre aux statisticiens pour regagner le terrain de la KDD occupé actuellement par les informaticiens.

A ce titre le futur de R est prometteur en visant à combler cette lacune. Le premier pas a été fait en mettant à disposition des interfaces avec des bases de données relationnelles.

Il y a plusieurs types de DBMS ou SGBD («Data Base Management Systems» ou «Système de Gestion de Base de Données»). Parfois l'utilisateur a le choix du DBMS. Plus souvent, il doit utiliser une base de données existante ce qui ne lui laisse pas le choix. La plupart des DBMS relationnels sont des systèmes du type client/serveur, et beaucoup permettent la transmission par TCP/IP. La plupart des DBMS viennent avec un moniteur, un client basé texte, certains ont des clients GUI et presque tous ont une API C ou C++. Notons que seul le DBMS peut accéder à la base de données par l'intermédiaire de commandes qui sont normalement introduites avec un dialecte de SQL («Structured Query Language» ou «Langage de Requête Structuré»). SQL est le principal langage informatique de création et de manipulation de bases de données relationnelles permettant tout aussi bien la définition que la manipulation et le contrôle d'une base de données relationnelles. Le but de cet article n'est pas de vous donner une introduction à SQL que vous trouverez sur de nombreux sites de la toile, dans le manuel et les HOWTOs de votre système de gestion de bases de données ou dans des livres.
Actuellement, il y a sur le "Comprehensive R Archive Network" [6] trois modules d'interface de R avec des DBMS :

- RMySQL pour MySQL;
- RmSQL pour MiniSQL;
- RPgSQL pour PostgreSQL;

utilisant des interfaces en C. Celles-ci et RODBC sont décrits dans le manuel «R Data Import/Export» [7] et illustrés dans [2]. L'option la plus portable est de loin RODBC et, à moins que vous puissiez choisir votre système de gestion de base de données, vous n'aurez probablement pas d'autre choix.

D'autres langages que R fournissent des interfaces de programmation bien définies pour des DBMS, tels que JDBC, Perl DBI et d'autres. R n'étant pas seulement un langage de programmation, mais également un outil pour l'analyse de données, il n'a pas seulement besoin d'une interface de programmation mais aussi d'une interface utilisateur avec DBMS.

Pour des informations concernant les futures directions de R et de ses interfaces avec des bases de données, je vous renvoie aux articles [8] et [9] et surtout au "Omegahat Project for Statistical Computing" [10]. C'est un projet collectif dont le but est de fournir une variété de logiciels libres pour des applications statistiques. Ce projet a débuté en 1998, avec des discussions entre les créateurs responsables de trois langages statistiques actuels (S, R, et LispStat). Le projet Omegahat développe également une collection de modules pour soutenir les nouvelles directions dans la programmation dans le langage S (comme mis en application dans R ou dans S-PLUS). Les modules illustrent comment communiquer entre R et d'autres langages et applications. Il existe des modules pour traduire en R du code écrit dans un autre langage ou pour faire appel à des fonctions R depuis un autre langage ou application. Ceci englobe aussi la possibilité d'inclure R dans un tel module ou vice versa.

Conclusion
Dans ce document j'ai brièvement introduit R (voir aussi [1]) et j'ai brièvement illustré les interfaces existantes de R avec des bases de données relationnelles comme étant un premier pas des logiciels statistiques modernes et libres, comme R, pour regagner le domaine du KDD (voir aussi [2]). De plus, n'oublions pas que R fournit une grande variété de techniques statistiques et graphiques et est fortement extensible. Le langage de S est souvent le véhicule de choix pour la recherche dans la méthodologie statistique et R fournit un itinéraire libre pour participer à cette activité.

Pour souligner ce propos, l'article [3] montre l'utilisation de la géostatistique en combinant R avec GRASS GIS.

Bibliographie
[1] Diego Kuonen & Valerie Chavez-Demoulin (2001). «R - un exemple du succès des modèles libres». Flash Informatique de l'EPFL, 2, 3-7. http://sawww.epfl.ch/SIC/SA/publications/FI01/fi-2-1/2-1-page3.html

[2] Diego Kuonen & Reinhard Furrer (2001). «Data mining avec R dans un monde libre». Flash Informatique de l'EPFL, Spécial Été, 45-50. http://sawww.epfl.ch/SIC/SA/publications/FI01/fi-sp-01/sp-1-page45.html

[3] Reinhard Furrer & Diego Kuonen (2001). «GRASS GIS et R : main dans la main dans un monde libre». Flash Informatique de l'EPFL, Spécial Été, 51-56. http://sawww.epfl.ch/SIC/SA/publications/FI01/fi-sp-01/sp-1-page51.html

[4] Ross Ihaka & Robert Gentleman (1996). «R: A Language for Data Analysis and Graphics». Journal of Computational and Graphical Statistics, 3, 299-314.

[5] Kurt Hornik (2001). «The R FAQ». http://cran.r-project.org/doc/FAQ/R-FAQ.html

[6] «The Comprehensive R Archive Network (CRAN)» http://cran.r-project.org http://cran.ch.r-project.org

[7] «The R Project for Statistical Computing»: http://www.r-project.org

[8] Brian D. Ripley (2001). «Using Databases with R». R News, 1, 18–20.

[9] Torsten Hothorn, David A. James & Brian D. Ripley (2001). «R/S Interfaces to Databases». In Proceedings of the 2nd International Workshop on Distributed Statistical Computing, March 15-17, Vienna, Austria: http://www.ci.tuwien.ac.at/Conferences/DSC-2001

[10] «The Omegahat Project for Statistical Computing»: http://www.omegahat.org

[11] Statoo Consulting, Statistical Consulting + Data Analysis Services, Lausanne: http://www.statoo.com

RÜCKBLICK KURS STATISTICAL DATA MINING

Die Schweizerische Gesellschaft für Statistik hat vom 3.-5. Oktober 2001 einen dreitägigen Kurs zum Thema «Statistical Data Mining» mit Prof. Brian Ripley von der Oxford University organisiert. Der Kurs stiess auf ein sehr grosses Interesse und war in kürzester Zeit ausgebucht.

Insgesamt 24 Teilnehmerinnen und Teilnehmer folgten dem sehr interessanten Kurs mit einem hervorragenden Dozenten im Ausbildungszentrum Unterhof in Diessenhofen. Die Teilnehmer waren Statistiker aus Versicherungen und Banken, statistische Beratungsunternehmen, Statistiker aus der Airline Industrie, aus der chemischen Industrie, von Hochschulen und auch aus der öffentlichen Statistik.

Der grosse Erfolg des Kurses und die positiven Rückmeldungen ermutigen den Vorstand der Gesellschaft, auch im nächsten Jahr einen weiteren Kurs anzubieten. Voraussichtlich wird im kommenden Herbst ein Kurs zu einem neuen Thema am gleichen Tagungsort stattfinden.

Caterina Savi