Data Mining

Adrien Brochot, Etienne Brodu

10 mars 2011

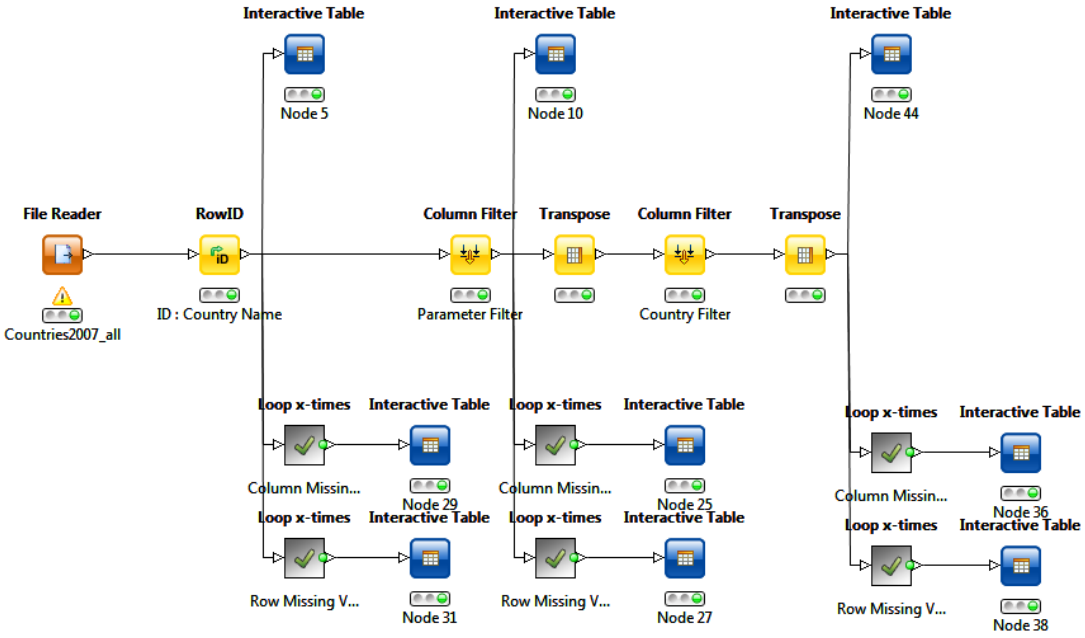
Table des matières

0.1 Introduction

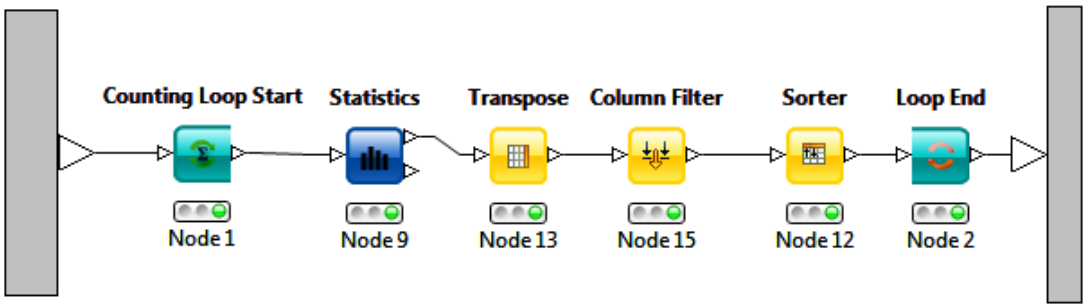
à partir d'un jeu de donnée fourni par the World Bank Group en 2007, nous avons essayé de déterminer certaines corrélations et de faire apparaitre des liens parmis ces données. Ces dernières contiennent des informations concernant 48 atributs sur 209 pays. Cependant certaines données sont manquantes, la première étape a donc été de trier les données pour ne récupérer que des données valides. Il aurait pu être intéressant d'essayer de retrouver certaines données manquantes, mais nous n'avons pas choisi d'effectuer cette étude. Durant notre étude, nous n'avons travaillé que sur le fichier countries_2007_all.csv. Nous avons préféré supprimer uniquement les données manquante à notre étude plutôt que de trier un jeu de données déjà réduit.

0.2 Tri des données

Nous savions que, quelque soit l'étude choisie, il faudrais éliminer un certains nombre d'attributs et un certains nombre de pays pour lesquelles les données manquent. Pour effectuer cette opération, nous avons mis en place une chaîne de selection permettant de trier précisément, par selection, les attributs puis les pays à inclure dans l'étude. Une solution alternative aurait été de trier uniquement les attributs interessant, puis supprimer tous les pays pour lesquelles des données manquent, mais afin de pouvoir vérifier quels pays doivent être enlevés, la première solution à été préférée. Un accent à été mis sur la visualisation des données manquantes. Grâce à des selecteurs, nous avons mis en place une chaîne permettant de connaître précisement le nombre de pays pour lesquelles manque un attribut, et le nombre d'attributs qui manquent à un pays. Grâce à ces selecteurs, nous faisions les tris necessaires afin d'obtenir uniquement un ensemble de données sans données manquantes.
La figurre suivante correspond au diagramme nous permettant de trier les jeux de données afin de sélectionner les colonnes à étudier et de supprimer éléments contenant des valeurs manquantes dans les colonnes étudiées.



Dans ce diagramme, nous commençons par lire le jeu de données complet à l'aide d'un **File Reader**. Le bloc **RowID** nous permet ensuite de définir la colonne des noms de pays comme indexe du jeu de données. Nous avons ensuite utilisé à deux reprises le composant **Loop x-times** comme simple conteneur de blocs : afin de simplifier le diagramme général, nous avons encapsulé dans ce bloc les éléments suivants :



Nous n'utilisons ici le composant **Loop x-times** que par soucis de lisibilité du diagramme général, nous avons fixé le nombre d'itérations à 1. Cette fonction permet de compter le nombre de valeurs manquantes par colonne. Par exemple, on peut voir ci-dessous le nombre de valeurs manquantes par colonne dans le jeu de donnée de base :

Row ID	No. missings
Roads, paved (% of total roads)	206
Malnutrition prevalence, weight for age (% of children...	205
Income share held by lowest 20%	203
Births attended by skilled health staff (% of total)	194
Contraceptive prevalence (% of women ages 15-49)	193
Cash surplus/deficit (% of GDP)	121
Revenue, excluding grants (% of GDP)	120
Total debt service (% of exports of goods, services a...	115
Market capitalization of listed companies (% of GDP)	110

Nous utilisont alors un élément **column Filter** afin de supprimer les colonnes inutiles à notre étude.
Le même principe est appliqué à pour la suppression des pays dont les valeurs dans les colonnes étudiées sont manquantes. Nous commençons par tranposer la table pour avoir les pays en colonnes et nous filtrons alors les colonnes contenant des valeurs manquantes.

0.3 Visualisation, corrélation, choix des dimensions,

Nous avons fais plusieurs comparaison, afin d'essayer d'établir des corrélation avec le pourcentage d'utilisation d'internet.

- La première comparaison à été faite avec le GDP et le GNI.
- La seconde comparaison à été faite avec le pourcentage de souscription à un forfait mobile.
- La troisiéme comparaison à été faite avec le temps requis pour démarrer une entreprise.

0.3.1 Normalisation

Afin d'avoir des distances significatives, pour établir des cluster cohérent, les données sont normalisé. Si la normalisation n'est pas faite un des attributs à un domaine bien plus grand que l'autre, ce qui en résulte sur des clusters qui semblent ne pas dépendre de l'attributs qui à le domaine le plus grand. En effet la distance etant bien plus grand sur une dimension que sur l'autre, le nuage de point peut être imaginé comme une longue bandelette, et Knime découpe les cluster selon une seule dimension. La normalisation permet donc, sans modifier la répartition des points, d'avoir des calculs de distance plus efficaces, et donc des clusters cohérents.

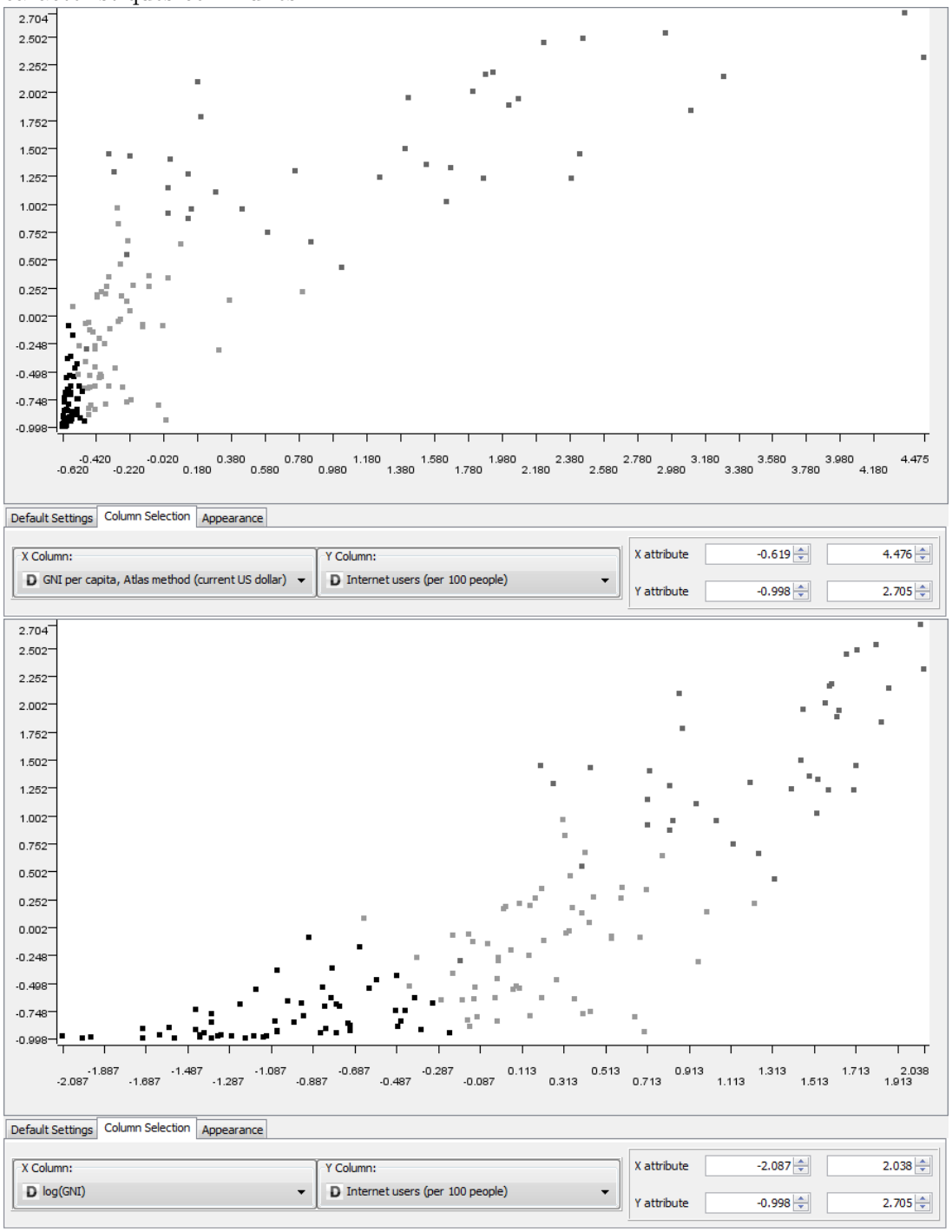
0.3.2 Echelle logarithmique

Dans le cas du GDP et du GNI, les données sont réparties de telles façon qu'un grand nombre de pays se trouve avoir des valeurs trés faibles, tandis que peu ont un GDP ou un GNI trés fort. La répartition selon cette dimension semblait logarithmique, nous avons donc essayé d'utiliser

une echelle logarithmique afin d'améliorer la visualisation des données. Grâce à cette nouvelle répartition, le nuage de point était réparti de manière bien plus homogéne.

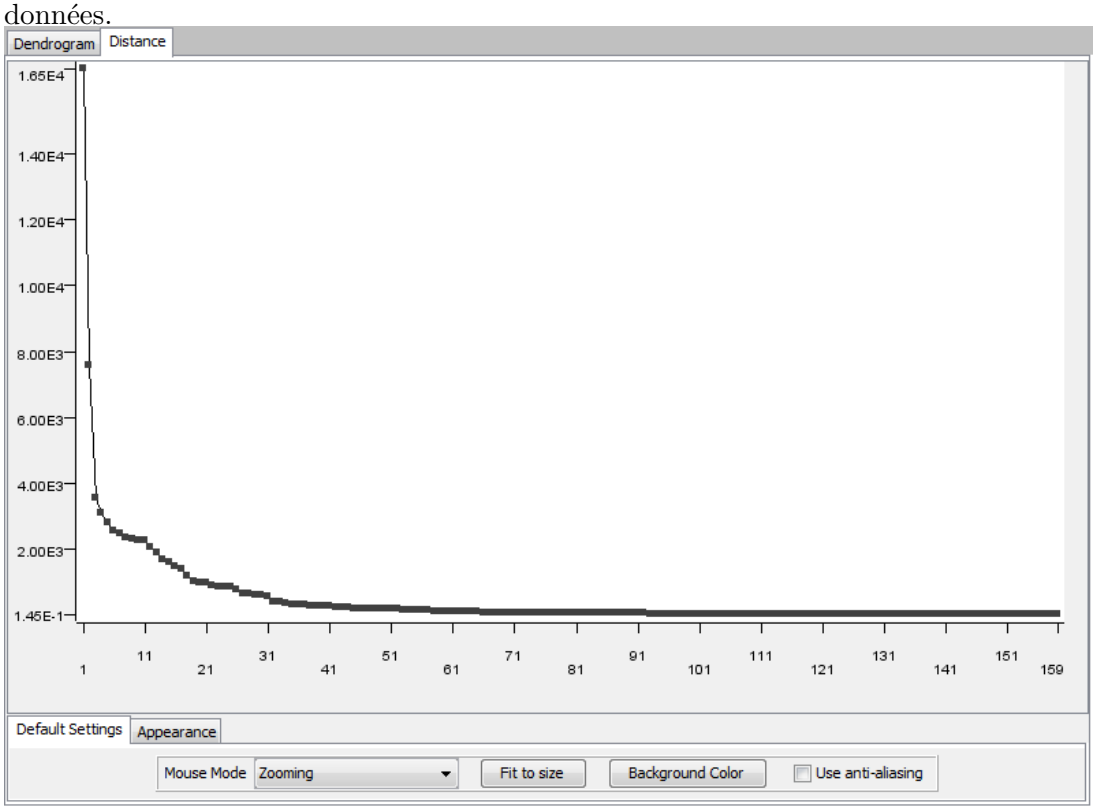
0.3.3 Internet Utilisation - GDPGNI

La comparaison entre GDP et GNI n'as pas permis de déceler de correlation évidente, à part des conclusions déjà connues concernant les extremes : un pays avec un trés faible GDP n'auras pas une forte utilisation d'internet, de même, un pays avec un fort GDP auras forcément une certaine utilisation d'internet. En revanche, on peut établir des groupes de pays ayant des caractéristiques communes :



Les résultats avec le GDP étant similaires mais moins visibles, nous ne présentons ici que ceux du GNI. La vue suivante représente la vue du **Hierarchical Clustering**. Cette vue nous a permis de définir le nombre de clusters à créer. Nous nous sommes limités à 3 clusters, bien que cette valeur soit située avant le point d'inflexion de la courbe de distance car la distance

entre 3 et 4 clusters était déjà trop faible et ne justifiait pas la création d'un nouveau groupe de données.



0.3.4 Internet UTilisation - Mobile subscription

Cette c

0.4 Cluster non guidé

0.5 Cluster guidé et prédiction appartenance