Fouille de données

Gaëtan Bloch & *Maxime* Gaudin *février 2011*

Table des matières

1 Description des données

Le jeu de données provient du site `http://worldbank.org`. Il concerne 210 pays et fournis des informations comme :

- Le pourcentage du Revenue National Brute (RNB) généré par l'agriculture, l'industrie ou les services.
- La quantité de produits importés et exportés.
- Des informations démographiques et géographiques comme la densité et l'age de la population ou la surface du pays.
- Des informations sanitaires comme la quantité de personnes infectées par le Virus d'immunodéficience Humaine (VIH).
- Des centaines d'autres indicateurs.

Par ailleurs, ce jeu de donnée semble être un très bon entrainement pour nous exercer au *datamining*. En effet, en plus d'exhiber des données très concrètes pour nous, il est évident que des corrélations sont présentes. Enfin, il pourra être très intéressant de voir ces corrélations d'abord à l'échelle du monde, puis à l'échelle de l'Europe par exemple.

2 Méthode de travail

Cette partie va tenter de décrire très précisément la démarche de travail que nous avons établi au début de ce TP, notamment pendant les deux séances préparatoires. En plus de vous indiquer quelles sont les étapes que nous avons suivi pour arriver à nos résultats, elles constituerons une bases de travail solide que nous pourrons améliorer tout au long des deux séances de *datamining*.

2.1 Filtrage des données

Contrairement à ce que nous avions cru avant de commencer, il faut très souvent élaguer son *dataset* afin de donner plus de sens aux valeurs.

2.1.1 Réduction des dimensions

Une des problématique majeure du *datamining* s'appelle la *malédiction de la dimensionalité*. Cette « malédiction » provient du fait que plus le nombre de dimensions du jeux de donnée (*i.e.* le nombre de colonnes) croit, moins la notions de distance n'a de sens. En effet, les distances ont tendance à se réduire proportionnelement avec le nombre de dimensions. Or, les algorithmes de *clustering* utilisent cette distance comme critère principal de choix à l'appartenance à tel ou tel *cluster*.

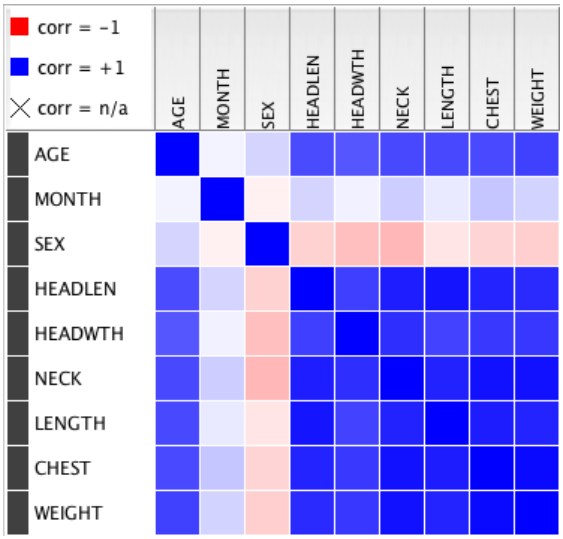
Voyons alors comment nous pouvons, dans le cadre d'un jeu de donnée contenant beaucoup de colonne, palier à ce défi technique.

Choix des colonnes

Dans un premier temps, il est fondamental de sélectionner les colonnes qui nous intéressent. Dans le cadre d'une étude de *datamining* réalisée par des professionnels, l'expérience dans le domaine analysé joue beaucoup et permettra de faire un choix cohérent et judicieux des dimensions à conserver. Cependant, il est fréquent que les données ne soient pas dans le domaine d'expérience du chercheur ou pire, très abstraites. Dans ce cas, il faut faire appel à d'autre technique.

Correlation

Il peut être par exemple très intéressant d'utiliser une matrice de correlation afin de mettre en valeurs les colonnes qui pourraient nous êtres utiles, et surtout celles qui nous sont inutiles.



La lecture de cette matrice est très simple. Plus les couleurs (rouge ou bleu) sont foncée, plus la corrélation, *i.e.* la facilité à déduire une colonne à partir d'une autre, entre les colonnes est forte. Il est évident que les diagonales soit de la couleurs la plus foncée possible car une colonne est toujours corrélée avec elle même.

Dans cet exemple, il est facile de déduire de la matrice que le sexe et le mois de mesure n'ont finalement pas beaucoup d'impact sur les autres paramètres. Nous pouvons les ignorer dans notre *clustering*.

PCA

La Primary Component Analysis (PCA) est une méthode de réduction automatique des dimension. Elle a l'avantage de préserver *mathématiquement* le plus d'information possible. Pour cela, elle essaie de chercher la meilleur combinaison linaire des différents colonnes afin d'atteindre, au choix *un nombre fixé de dimension* ou *une conservation de X% de l'information*. Bien que pratique, elle est (tout du moins à notre niveau d'expertise en *datamining*) relativement dangereuse car les combinaisons linéaires qui résulte d'une PCA n'ont plus vraiment de sens. Aussi, il est assez compliqué d'obtenir la formule de sortie afin de l'interpréter.

2.1.2 Valeurs manquantes

Certains *dataset* contiennent des lignes n'exhibant pas toutes les colonnes. Ce problème anodin peut pourtant poser des problèmes et il faut adopter une des stratégies suivantes :

- Supprimer l'ensemble des lignes ou au moins une valeur manque : C'est la solution la plus propre mais le jeu de donnée risque de se réduire drastiquement.
- Remplacer la valeur manquante : Soit par une moyenne des autres valeurs, soit par une valeur « label » (*e.g.* `missing`), soit même par une valeur aléatoire.

La meilleur solution reste toute de même de bien sélectionner ses colonnes, et de supprimer les lignes contenant une valeur manquante.

2.1.3 Outliers

Un *outlier* est une valeur atypique du jeu de donnée risquant de biaiser l'analyse que nous pourrions en faire. Cependant, elles peuvent aussi aider à traiter un problème dans un cas plus général. C'est pourquoi leur élimination (ou conversation) est problématique et fait appel aussi bien au bon sens, qu'a la compréhension des données ainsi qu'à l'expérience.

Recherche « à la main »

La première méthode de recherche a l'avantage d'être simple. Elle consiste simplement à effectuer une recherche visuelle des *outliers* via un *scatter plot* pour les jeux de données de dimension inférieur à 2 et via un *parallel coordinates*. L'humain étant très fort pour détecter les valeurs extrêmes, pour peu que l'opérateur a compris les données qu'il maniait et qu'il sait lire correctement un graphique, les *outliers* devraient être assez simple à exclure. Il pourra d'ailleurs utiliser les fonction d'*HiLighting* proposées par le logiciel.

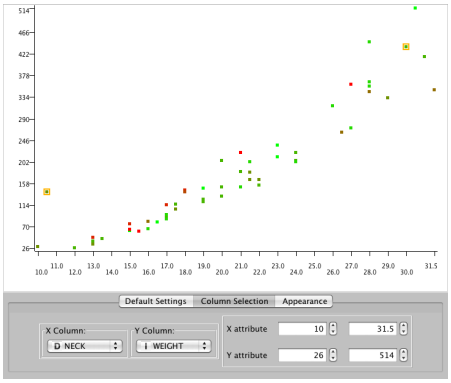


Figure 1 – Un exemple de diagramme « Scatter plot »

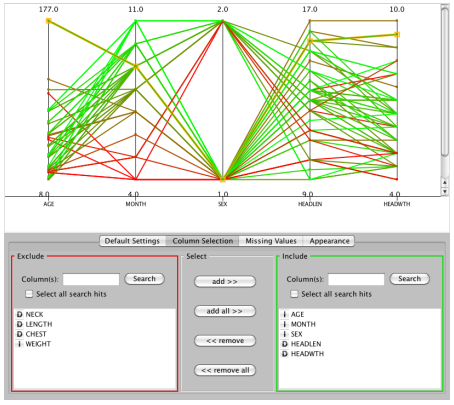


Figure 2 – Un exemple de diagramme « ParallelCoordinates »

Clustering hiérarchique

Une des méthodes efficace et qui me semble un peu plus rigoureuse consiste à utiliser un *clustering* hiérarchique afin de déterminer les *outliers*.

La méthode est très simple. En effet, il suffit de regarder quel point ou petit groupe de point s'est « collé » en dernier. Sur la figure 4, il est évident que le point le plus a gauche est un *outlier*. En effet, il s'est collé en dernier, et ce, à une distance presque égale au rayon du cluster contenant tous les autres points.

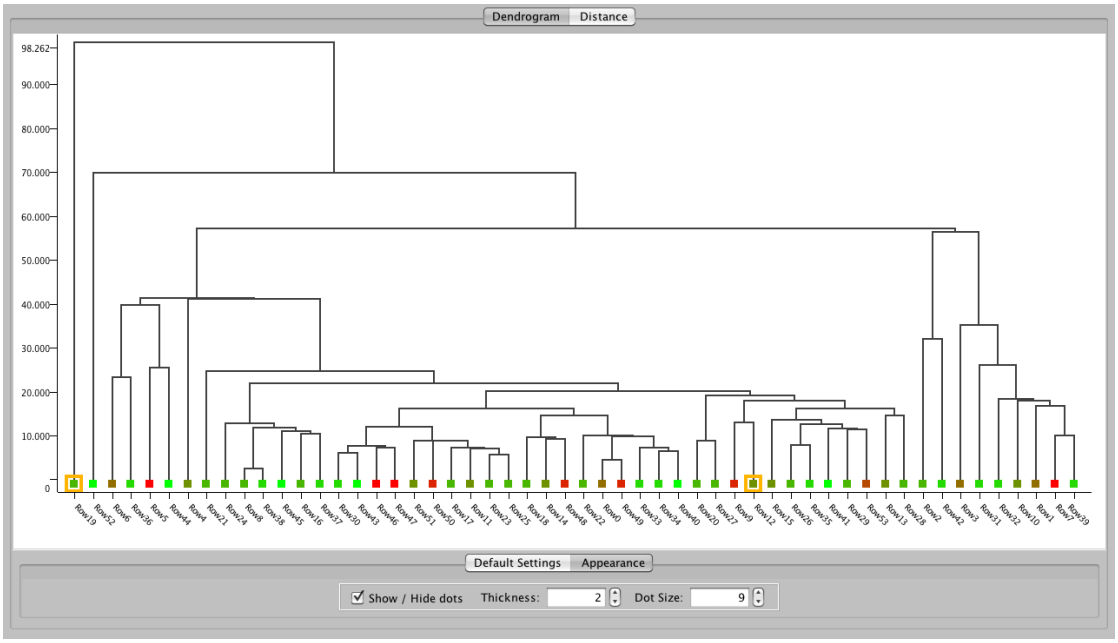


Figure 3 – Exemple de dendogramme extrait d'un *clustering* hiérarchique

Box plot

C'est encore un moyen très rapide de démasquer les outliers car ils font figurer sur les même diagramme, la médian, les quartiles et les décile. Nous avons donc toutes les données statistiques nécessaire afin d'évaluer l'appartenance d'une point à notre futur jeu de données.

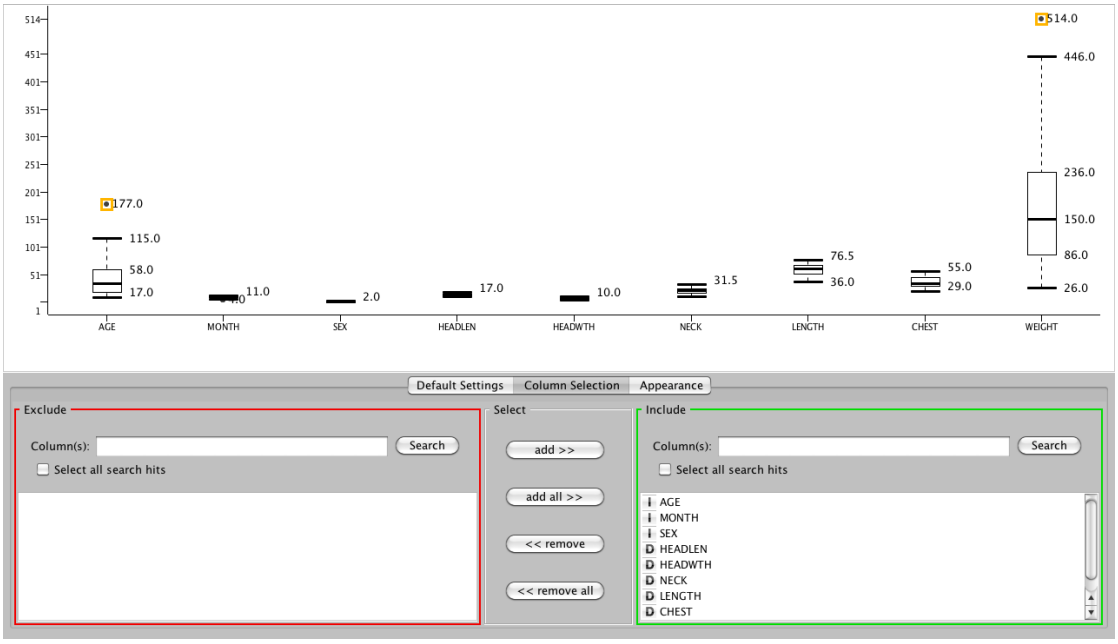


FIGURE 4 – Exemple de boîte à moustache

Méthodes statistiques

Une des dernières méthodes (que nous n'utilisons pas sur Knime) est une approche statistique des données. Il est prouvé que de nombreux phénomènes suivent une loi normale. De plus, nous connaissons très bien cette loi et des mathématiciens ont pu modéliser l'atypique afin de fournir des méthode automatique d'exclusion des *outliers*. Nous ne rentrerons pas plus dans le détail mais parmi ces méthodes, nous pouvons citer le critère de PEIRCE, de CHAUVENET ou encore le test de GRUBB.

2.2 Détermination du nombre de clusters

Intro ou le nombre est connu

Clustering hiérarchique

SINGLE : Bras

COMPLETE et AVERAGE : Paquets, et mieux adapté à *K-Means*

2.3 Clustering

Normalisation

Distance : Utilisation des gaps plus que de la dérivé.

2.3.1 Hierarchical Clustering

2.3.2 K-Mean

2.3.3 Fuzzy C-Mean

Faire la liste et les spécificité des différentes méthodes de *clustering*

3 Secteurs d'activité des différents pays

Ajouter le nom de la colonne correspondante

Pour cette première étude de cas, nous avons choisi d'utiliser les données afin de caractériser l'activité d'un pays en fonction de :

- La taille du pays
- La population
- Le Produit Intérieur Brute (PIB)
- Le RNB
- Les exports industriels

Ajouter l'environnement subimages pour une meilleur intégration

Notre démarche dans cette analyse est simple : Après avoir établi un *clustering* sur le pourcentage du PIB dû à l'agriculture, l'industrie et les services, nous avons créer un arbre de décision prenant pour critère de sélection les données citées ci-dessus.

3.1 Sélection des données

Après avoir sélectionné les colonnes qui nous intéressent (*i.e.* `Agriculture, value added (% of GDP)`, `Industry, value added (% of GDP)` et `Service, value added (% of GDP)`), nous obtenons la matrice suivante :

boite à chaussure ? ?

Le diagramme de est :

Enfin le diagramme *parallel coordinates* :

Guinée : 95% de son PIB est industriel

3.2 Clustering

3.2.1 Clustering hiérarchique

Le choix du type de *clustering* hiérarchique est, *a priori* délicat. Voici les résultats obtenus avec un *clustering* hiérarchique :

SINGLE

COMPLETE

3.2.2 Clustering K-Means

Inclure le tableau de vérification de l'entropie

Vérification de la stabilité du *clustering*

3.3 Interprétation

3.3.1 Arbre de décision

4 Création d'un module de vérification du K-Mean

Liste des accronymes

PCA

Primary Component Analysis [Page(s) 3]

PIB

Produit Intérieur Brute [Page(s) 6]

RNB

Revenue National Brute [Page(s) 1, 6]

VIH

Virus d'immunodéficience Humaine [Page(s) 1]

Fin