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` et fournis, à propos de 210 pays des données comme la population totale, la surface, le Produit Intérieur Brute (PIB) et sa croissance ou encore le nombre de malade atteint du Virus d'immunodéficience Humaine (VIH).

2 Méthode de travail

2.1 Filtrage des données

Normalisation

Valeurs manquantes

Outliers

Correlation

2.2 Réduction des dimensions

Choix des colonnes

PCA

2.3 Détermination du nombre de clusters

***Clustering* hiérarchique**

2.4 Clustering

Elle est parfois inutile (dans le cas de valeur en pourcentages *e.g.*

Il faut les chercher en 1D, 2D, et plus (parallel coordinates

Utilisation du clustering hiérarchique pour les identifier : Repérer les derniers collé

Utilisation des box plots

Intro ou le nombre est connu

SINGLE : Bras

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

Distance : Utilisation des gaps

2.4.1 Hierarchical Clustering

2.4.2 K-Mean

2.4.3 Fuzzy C-Mean

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 PIB
- Le À décrire (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 :

Guinée : 95% de son PIB est industriel

Enfin le diagramme *parallel coordinates* :

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