

Table des matières

1	Description des données	1
2	Méthode de travail	1
2.1	Filtrage des données	1
2.2	Réduction des dimensions	2
2.3	Détermination du nombre de <i>clusters</i>	2
2.4	<i>Clustering</i>	2
2.4.1	<i>Hierarchical Clustering</i>	2
2.4.2	<i>K-Mean</i>	2
2.4.3	<i>Fuzzy C-Mean</i>	2
3	Secteurs d'activité des différents pays	3
3.1	Sélection des données	3
3.2	<i>Clustering</i>	3
3.2.1	<i>Clustering</i> hiérarchique	3
3.2.2	<i>Clustering K-Means</i>	3
3.3	Interprétation	3
3.3.1	Arbre de décision	3
4	Création d'un module de vérification du <i>K-Mean</i>	4

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 **Produit Intérieur Brute (PIB)** généré par l'agriculture, l'industrie ou les services.
- La quantité de produits importés et exporté.
- 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 infectés par le **Virus d'immunodéficience Humaine (VIH)**.
- Des centaines d'autres indicateurs.

Par ailleurs, ce jeu de donnée semble être un très bon entraînement 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 les corrélations à l'échelle du monde dans un premier temps, puis à l'échelle de l'Europe par exemple.

2 Méthode de travail

2.1 Filtrage des données

Normalisation

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

Valeurs manquantes

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

Outliers

Correlation

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

2.2 Réduction des dimensions

Choix des colonnes

Utilisation des box plots

PCA

2.3 Détermination du nombre de clusters

Intro ou le nombre est connu

Clustering hiérarchique

2.4 Clustering

SINGLE : Bras

2.4.1 Hierarchical Clustering

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

2.4.2 K-Mean

Distance : Utilisation des gaps plus que de la dérivé.

2.4.3 Fuzzy C-Mean

Faire la liste et les spécificité des différentes méthodes de clustering

3 Secteurs d'activité des différents pays

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

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 créé 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 :

Le diagramme de est :

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

Vérification de la stabilité du *clustering*

3.3 Interprétation

3.3.1 Arbre de décision

Ajouter le nom de la colonne correspondante

Ajouter l'environnement subimages pour une meilleur intégration

boite à chaussure ??

Guinée : 95% de son PIB est industriel

Inclure le tableau de vérification de l'entropie

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

Liste des accronymes

PIB

Produit Intérieur Brute [Page(s) [1](#), [3](#)]

RNB

À décrire [Page(s) [3](#)]

VIH

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

Fin

