Master 1 Informatique
Université des Sciences et Technologies Lille 1

Business Intelligence

Rapport TP1 Clustering

Saint-Omer Anne-Sophie & Poux Delphine

Segmentation de la base de données « Square 1»

1 La distance utilisée est la distance Euclidienne car nous devons connaitre la distance entre des points sur un graphe à deux dimensions.

2 En ce qui concerne la moyenne, nous prenons la moyenne arithmétique.

3 Voir les sources

4 Pour utiliser la base de données, on exécute le main MainSquare. Nous avons également vérifié notre code avec MainPoint

5 Les résultats obtenus sont :

   a WC = 2494.2039961475375

   b BC = 67.49984438687054

   c Rapport BC/WC = 0.027062679913563002

6 En essayant d'autres choix de K, on a WC qui augmente considérablement ou BC qui baisse beaucoup. Les distances de Manhattan et de Chebyshev donnent des résultats similaires, ils semblent tous corrects pour le jeu de données square.

7 Nous utilisons le graphe suivant avec gnuplot :

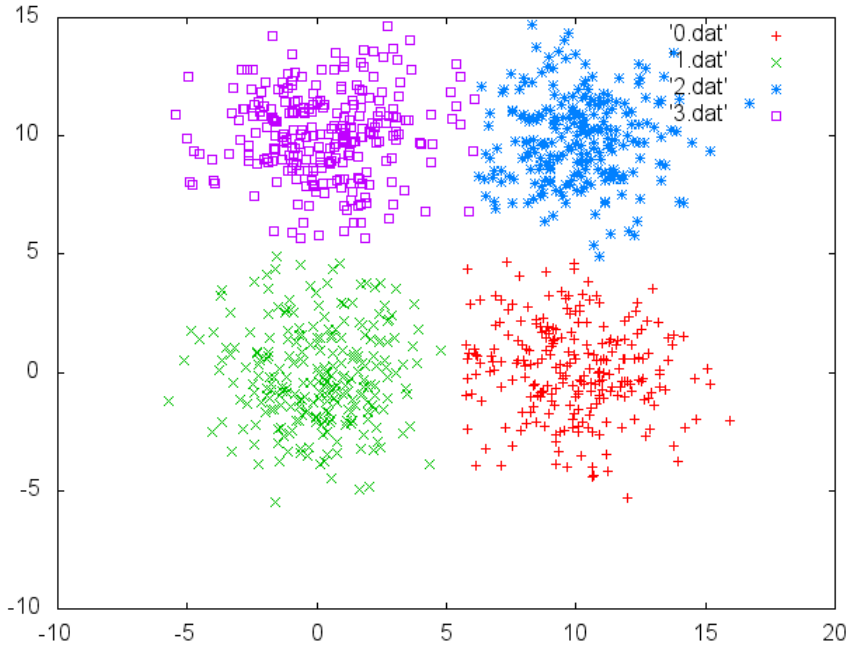

FIGURE 1 – Répartition des clusters Mainsquare

8 On peut constater sur ce graphe que 4 clusters ont été trouvés. Ces 4 clusters regroupent les 4 "parties" où les points sont les plus proches.Le résultat obtenu nous parait correct, il y a bien 4 clusters correspondants aux 4 "parties" sur le graphe.

Segmentation de la base de données "Iris"

1 La distance utilisée est la distance euclidienne pour les 4 premiers attributs car ce sont des attributs numériques.

2 Pour la moyenne, nous pouvons garder la même que pour Square.

3 voir sources

4 Afin d'utiliser la base de données Iris, nous exécutons MainIris.

5 Les résultats sont :

a WC = 97.34621969415679

b BC = 10.122894049846408

c Ratio BC/WC = 0.10398856865372488

d Le tableau de contingence est :

	Iris Setosa	Iris Versicolor	Iris Virginica	Total
Cluster 0	0	3	36	39
Cluster 1	50	0	0	50
Cluster 2	0	47	14	61
Total	50	50	50	150

On peut donc associer le cluster 0 à la classe Iris Virginica, le cluster 1 à Iris Setosa et le cluster 2 à Iris Versicolor.

6 En augmentant K, WC augmente beaucoup.

Lorsqu'on diminue K, WC augmente et BC diminue beaucoup.

Avec la distance de Manhattan, WC augmente et BC diminue.

Tous ces cas nous donnent un ratio moins intéressant.

La distance de Chebyshev est à peu près équivalente à la distance euclidienne.

7 Le résultat obtenu est :

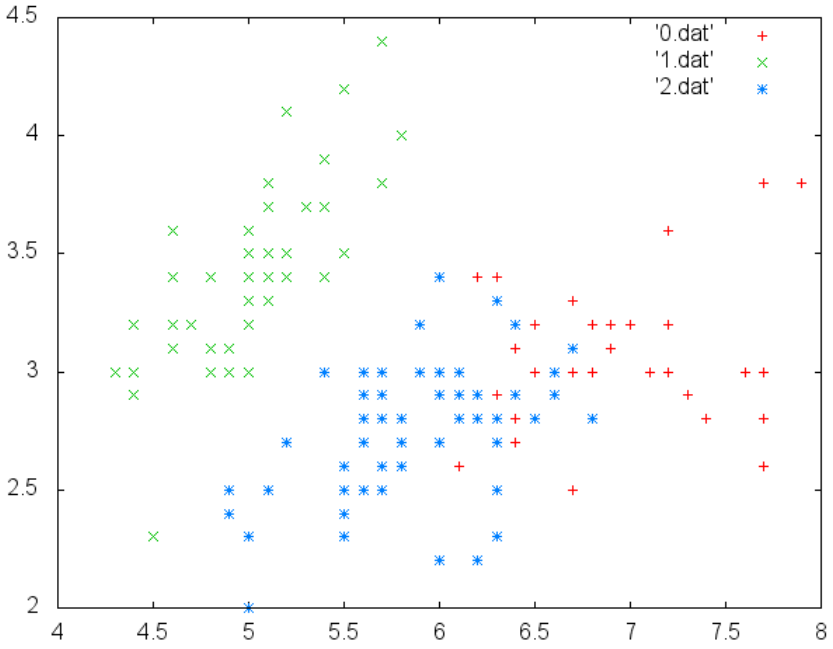

FIGURE 2 – Répartition des clusters Iris

8 La répartition a un taux d'erreur supérieur au jeu de données Square1. En effet, Square1 paraissait avoir très peu d'erreur (qu'on ne peut qualifier car nous ne pouvons pas construire le tableau de contingence étant donné qu'il n'y a pas de classe) alors que Iris possède $3+14 = 17$ erreurs. Soit un taux d'erreur de $\frac{17}{150} * 100 = 11,33\%$ Cependant, la répartition parait tout de même assez correcte.

Segmentation de la base de données "Titanic"

1 Pour les attributs age, sex et survived, nous avons deux valeurs possibles : 1 et 0. On peut donc prendre comme distance :

$$dist(v1, v2) = |v1 - v2|$$

Ce qui nous donnera 1 si v1 = v2, 0 sinon.

2 Pour la moyenne, nous pouvons utiliser la moyenne arithmétique. Les valeurs des attributs sont 0 ou 1. La moyenne sera donc forcément comprise entre 0 et 1.

Par exemple si la moyenne pour l'attribut sexe est plus proche de 0, cela signifie qu'il y a plus de femmes dans le cluster.

Dans ce cas, la distance entre la moyenne et une donnée sera plus proche si c'est une femme : si moy < 0.5 alors

$$|moy - 0| < 0.5$$

et

$$|moy - 1| > 0.5$$

donc

$$|moy - 0| < |moy - 1| \leftrightarrow dist(moy, Dfemme) < dist(moy, Dhomme)$$