Algorithme k-Means à deux modes

1. Introduction

Clustering peut être considérée comme une des pierres angulaires de la classification. Soit un ensemble de données dans les deux sens typique bimode des répondants par des variables. Souvent, les algorithmes de regroupement sont appliqués à un seul mode de la matrice de données, qui peut être fait de manière hiérarchique ou non hiérarchique (J. van Rosmalen 2009).

Une forme relativement nouvelle de regroupement est grappe à deux modes. Dans groupement bi-mode, les lignes et les colonnes d'une matrice de données à deux modes sont attribués à des clusters. Chaque rangée d'une matrice de données à deux modes est associée à une ou plusieurs grappes de rangée, et chaque colonne à un ou plusieurs groupes de colonnes. Les éléments de la matrice de données qui sont dans le même groupe de rangée et dans le même groupe de colonnes doivent être proches (J. van Rosmalen 2009).

Dans cet article, nous nous concentrons sur la mise en œuvre bimode algorithme de partitionnement k-means, qui est une version étendue de k-means classiques.

1. Determination de l'algorithme

|  |  |
| --- | --- |
| Notation | Description |
| X n×m = x[I][j] | Matrice de données à deux modes de n lignes et m colonnes. |
| P n×K = p[I][k] | Matrice d'appartenance de cluster des lignes avec K le nombre de clusters de lignes, p[i][k] = 1 si la rangée i appartient au cluster de lignes k et p[i][k] = 0 sinon. |
| Q m×L = q[j][l] | Matrice d'appartenance de cluster des colonnes avec L le nombre de clusters de colonnes, q[j][l] = 1 si la colonne j appartient au groupe de colonnes l et q[j][l] = 0 sinon. |
| V K×L = v[k][l] | Matrix with cluster centers for row cluster k and column cluster l. |
| E n×m = e[I][j] | Matrix with errors from cluster centers. |

, (1)

(2)

L'algorithme K-means (Hartigan 1975) est une des plus simples et moyens les plus rapides pour obtenir une bonne partition, ce qui explique sa popularité dans le regroupement d'un-mode. Cet algorithme peut facilement être étendu à gérer deux partitionnement de mode (Baier, Gaul, and Schader 1997; Vichi 2001). Le soi-disant algorithme à deux modes k-means vise à améliorer une partition initiale avec l'utilisation (1) et (2) et est composé des étapes suivantes (J. van Rosmalen 2009).

1. Choisissez le premier P et Q.
2. Répétez ce qui suit, jusqu'à ce qu'il n'y ait pas d'amélioration de f (P, Q, V) dans n'importe quelle étape.
   1. Mettez à jour V selon (1).
   2. Laissez-vous . Puis mettez à jour P selon

(3)

* 1. Mettez à jour V selon (1).
  2. Laissez-vous . Puis mettez à jour Q selon

(4)

L'algorithme k-means à deux modes converge toujours vers un minimum local, car la valeur du critère f (P, Q, V) ne peut pas augmenter à aucune étape. Toutefois, une ou plusieurs grappes peuvent devenir vides après l'étape 2b ou 2d. Cette situation est immédiatement corrigée en transférant l'objet de ligne ou de colonne avec la valeur la plus élevée de (J. van Rosmalen 2009).

2. Les expériences que nous avons réalisées sur les données simulées

Nous générons la matrice de données X dans chaque instance de problème en simulant P, Q, V et E, puis en utilisant (5) pour construire X. La génération de données simulées de cette façon est naturelle et présente l'avantage de l'existence d'une structure de cluster dans les données . En outre, les P, Q et V utilisés dans la génération de X peuvent donner une limite supérieure utile sur la valeur optimale de f (P, Q, V).

(5)

Un grand nombre de facteurs peuvent être variés dans une étude de simulation pour le regroupement en deux modes, telles que les valeurs de n, m, K et L, la taille, la répartition des erreurs et d'autres. Pour cette étude, nous limitons le nombre de facteurs. Nous ne modifions que les paramètres n, m, K, L et les écarts de déviation d'erreur (J. van Rosmalen 2009).

Nous utilisons un critère pour évaluer les résultats de l'étude de simulation, à savoir le critère de variance expliqué (variance accounted for - VAF) (6), qui est comparable à la mesure R2 utilisée dans l'analyse de régression. Le critère VAF est défini comme:

(6)

où . La valeur optimale de VAF varie de 0 à 1, et la maximisation de VAF correspond à la minimisation de f (P, Q, V).

Afin de montrer comment fonctionne l'algorithme, nous avons décidé de réaliser trois tests dont les résultats sont donnés ci-dessous. Les figures représentent les données regroupées par classes en ligne et en colonne par l’algorithme:

Figure 1: Représentation graphique des données regroupées avec des tailles n = m = 60, K= L = 3, erreur écart = 0.5, VAF = 0.648

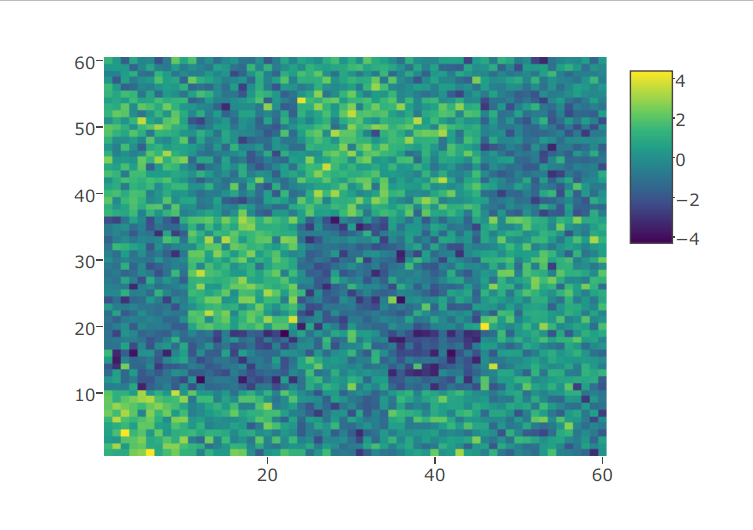
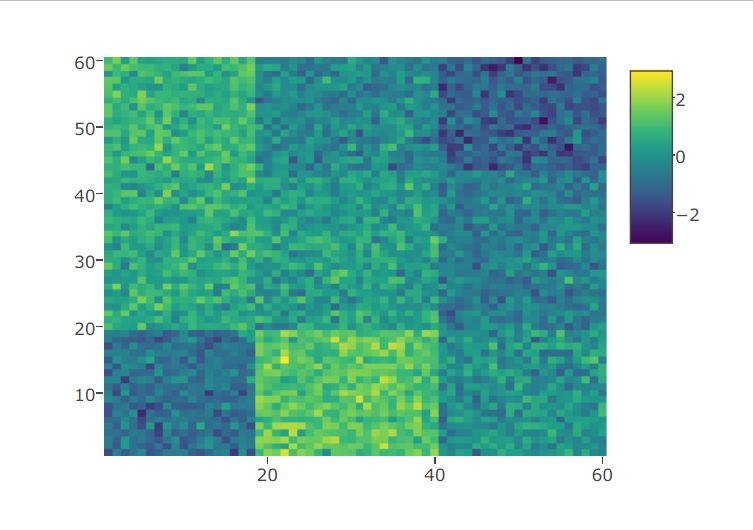


Figure 2: Représentation graphique des données regroupées avec des tailles n = m = 60, K = L = 5, erreur écart = 1, VAF = 0.468

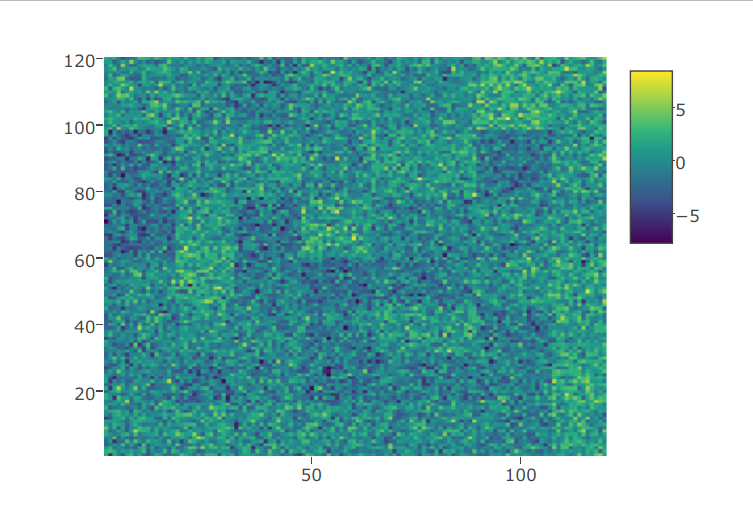


Figure 3: Représentation graphique des données regroupées avec des tailles n = m = 120, K = L = 7, erreur écart 2, VAF = 0.174

1. Les conclusions

Dans le cadre de ce travail d'étude, on a mis en place un algorithme k-means en deux modes, qui est une version étendue de l'algorithme classique de clustering k-means. Afin de montrer les résultats obtenus, on a utilisé des cartes de chaleur qui représentent graphiquement l'état final de la matrice en grappe. En outre, différents critères ont été impliqués, montrant comment l'algorithme se comporte avec l'une ou l'autre configuration.

Les références

BAIER, D., GAUL, W., and SCHADER, M. (1997), “Two-Mode Overlapping Clustering with Applications in Simultaneous Benefit Segmentation and Market Structuring,” in Classification and Knowledge Organization, eds. R. Klar and O. Opitz, Heidelberg: Springer, pp. 557-566.

HARTIGAN, J. A. (1975), Clustering Algorithms, New York: John Wiley and Sons.

VAN ROSMALEN, J., GROENEN, P.J.F., TREJOS, J. et al. (2009), Optimization Strategies for Two-Mode Partitioning, Journal of Classification, 26:155-181

VICHI, M. (2001), “Double k-Means Clustering for Simultaneous Classification of Objects and Variables,” in Advances in Classification and Data Analysis. Studies in Classification, Data Analysis, and Knowledge Organization, eds, S. Borra, R. Rocchi, and M. Schader, Heidelberg: Springer, pp. 43-52.