Université de Versailles Saint Quentin en Yvelines - UFR des sciences

RAPPORT D'IN511

Projet de PAM

Réalisé par :
M Théo FRATCZAK
M^{me} Céline SAOUDI
M. Yassine MAZIZ

Version du
8 décembre 2021

1 Introduction

Le projet consiste à écrire un algorithme de PAM qui permet la recherche de k objets représentatifs dans un ensemble de données (appelés k centroïdes),il attribue ensuite chaque objet au centre le plus proche afin de créer des clusters. Le but est de minimiser la somme des dissemblances entre les objets d'un cluster et le centre du même cluster (centroïde). Il est connu pour être un algorithme plus robuste que celui des k-means puisqu'il est considéré comme moins sensible aux valeurs aberrantes.

2 L'Algorithme

Cet algorithme est programmé en langage C.

2.1 Lecture du fichier :

-Notre projet contient un fichier Dataset fourni qui contient 50 noms de students avec N objets

-Dans un premier temps on ouvre le fichier avec la fonction fopen()

-On lit Ensuite le fichier avec la fonction fscanf en respectant la bonne lecture des noms, les objets (courage, loyauté, sagesse, malice , maison), Les données collectées sont enregistrées dans une structure représentant un étudiant (voir le fichier lib.h)

2.2 initialisation des clusters :

Le programme va choisir aléatoirement k objets (k est donné par l'utilisateur) qui seront chacun représentant d'un cluster (k clusters). L'algorithme va ensuite affecter tout objet non représentatif dans le cluster à l'objet représentatif le plus proche de celui-ci(En terme de distance).

2.3 Le calcul de coût S et E :

int cost(struct Cluster* clusters, struct Student* Students_data,struct Student* Students_rep, int k) est une fonction qui :
Pour chaque cluster et pour chaque objet o de la DATA celle-ci calcule le cout du partitionnement si l'objet o était représentant du cluster r. Au final, l'algorithme choisit le meilleur changement (r,o), si celui-ci améliore le cout du partitionnement, il est effectué.

2.4 L'algorithme

En premier lieu l'algorithme lit le fichier (Lecture du fichier)
Initialise les clusters (Initialisation des clusters)
Tant que des changements d'objet représentatif sont effectué, L'algorithme effectue une boucle sur la fonction cost().
(Voir dans le fichier main.c).

3 Complexité de l'algorithme

La complexité de l'algorithme PAM est de : (k*n)*2(k*n)

4 K-MEANS

En utilisant R pour les k-means on obtient ces résultats :

— > t=read.csv2(file.choose(), header= T,row.names = 1)
— > t
— > B=t[c(1 :4)]
— > kmeans(B, centers=4, nstart=10)
— > print(kmeans)

K-means clustering with 4 clusters of sizes 15, 12, 14, 9

Cluster means :

— Courage LoyautÃ. Sagesse Malice
— 1 9.533333 5.000000 3.866667 6.133333
— 2 5.166667 6.750000 5.333333 9.250000
— 3 4.214286 7.071429 8.071429 4.500000
— 4 9.111111 7.111111 9.111111 8.111111

Clustering vector :

Adrian *Andrew* Angelina *Anthony* Arthur
4 1 1 3 1
Bellatrix *Bole* Colin *Cormac* Dean
4 2 1 1 1
Demelza *Derrick* Eddie *Ernie* Euan
1 3 3 3 1
Gilderoy *Gregory* Hannah *Harper* Jimmy
4 2 1 2 4
Justin *Katie* Lavande *Lee* Luna
2 1 4 1 3
Marcus *Marietta* Michael *Milicent* Mimi
2 4 3 1 2
Montague *Neville* Norbert *Nymphadora* Padma
2 1 3 2 2
Paenny *Pansy* Parvati *Pomona* Quirinus
3 3 1 2 3
Roger *Romilda* Saemus *Sirius* Susan
4 1 3 4 3
Sean *Ted* Tarence *Terry* Vincent
3 3 2 4 2

— Within cluster sum of squares by cluster :
181.20000 170.83333 209.71429 89.55556
(between$_S S$/*total*$_S S$ = 52.2%)

Available components :
[1] "cluster" "centers" "totss" "withinss"

[5] "tot.withinss" "betweenss" "size" "iter"

"ifault"

5 La comparison des résultats de l'algorithme de k-means et l'algorithme de PAM

En posant k=4 par exemple :
D'une part on obtient 4 clusters qui minimisent la somme des distances de chaque sommet au centre de son cluster dans l'algorithme des k-means.
Et dans l'algorithme de PAM on obtient 4 clusters qui minimisent le coût.
D'autre part on remarque que chaque cluster dans k-means correspond a un cluster dans l'algorithme de PAM tel que :

— dans l'algorithme de PAM on a(Marietta - Bellatrix - Gilderoy- Jimmy- Lavande-Gryffondor - Roger - Sirius-Gryffondor) correspondent au cluster 4 dans l'algorithme des k-means avec les students(Bellatrix, gilderoy,Roger,Marietta,Lavande,Sirius,Jimmy,Adrien,Terry)

— (Padma - Adrian- Angelina - Bole - Derrick - Gregory - Harper - Justin- Marcus-Mimi- Montague - Nymphadora - Pansy -Romilda - Susan – Tarence-Vincent) correspondent au cluster 2 dans l'algorithme des k-means avec les students (Justin,Marcus,Montagne,Bole,Grerory,Tarence,Susan,Mimi,Harpper,Nymphadora,Padma,Pamona,Vincent)

— (Ted - Anthony - Eddie - Ernie - Hannah - Luna – Norbert - Paenny - Pomona-Quirinus-Saemus- Susan) correspondent au cluster 3 dans l'algorithme des k-means avec les students(Paenny,Sean,Derrick,Pansy,Ted,Eddie,Michael,Saemus,AnThony,Ernie,Luna, Quirinus, Susan,Nelbert)

— (Milicent - Andrew - Arthur - Colin - Cormac - Dean - Demelza - Euan- Katie-Lee- Michael - Neville-Parvati - Terry) correspondent au cluster 1 dans l'algorithme des k-means avec les students(Damelza,Andrew,Katie,Neville,Romilda,Parvati,Clin,Angelina,Hannah, Milicent ,Cormar,Arthur,Dean,Euan,Lee).