IF 3270 Pembelajaran Mesin

Implementasi K-Means dan Agglomerative Clustering

Disusun oleh:

Muhammad Akmal / 13517028
Haris Salman Al - Ghifary / 13517052
Hafidh Rendyanto / 13517061
Muhammad Nurdin Husen / 13517112

Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung
Bandung

2020

A. Penjelasan Implementasi

a. Implementasi K-Means Clustering

Pada bagian ini, kami membagi pengerjaan K-Means Clustering menjadi 4 tahap, yaitu:

1. Memilih secara acak centroid awal
   Pada tahap ini, dilakukan inisialisasi awal yaitu pemilihan secara acak k buah centroid dari titik yang telah tersedia. Centroid-centroid akan dilakukan perhitungan dan mempengaruhi pemilihan centroid berikutnya

2. Menghitung nilai centroid selanjutnya
   Pada tahap ini, dilakukan penghitungan nilai centroid selanjutnya berdasarkan pengelompokan yang telah dilakukan sebelumnya. Caranya dengan melakukan penghitungan rata-rata x dan rata-rata y pada setiap clusternya, dan ditemukan titik centroid baru. Pada tahap awal sekali, tahap ini dilewati karena belum ada pengelompokan.

3. Menghitung jarak setiap titik dengan semua centroid
   Pada tahap ini, penghitungan dilakukan untuk setiap titik yang ada terhadap titik-titik centroid yang tersedia. Penghitungan jarak dilakukan dengan Euclidean distance yaitu

$$Euclidean\ distance\ (d) = \sqrt{\sum_{i=1}^{n}(x_{i2} - x_{i1})^2}$$

   dengan x_{ij} adalah nilai atribut i untuk titik j

4. Mengelompokkan setiap titik terhadap centroid
   Setelah dilakukan penghitungan jarak, maka untuk setiap titik dilakukan pengelompokan dengan cara memilih centroid yang memiliki jarak terdekat dari suatu titik dibanding centroid lainnya.
5. Pengecekan terhadap perubahan centroid
   Jika tedapat perubahan centroid, maka proses clustering akan kembali lagi pada tahap 2 sampai tidak ada lagi perubahan centroid jika dibandingkjan dengan nilai centroid yang sebelumnya.

b. Implementasi Agglomerative Clustering

Implementasi agglomerative clustering dilakukan dalam beberapa tahap berikut ini;

1. Iterasi seluruh kemungkinan pasangan cluster
   Pada tahap ini, kami mengiterasi seluruh kemungkinan pasangan cluster untuk mencari pasangan cluster dengan jarak terkecil. Fungsi jarak yang kami gunakan adalah fungsi jarak yang sesuai dengan pilihan user, yaitu salah satu dari keempat pilihan berikut {Single Linkage, Complete Linkage, Average Linkage, Average-Group Linkage}. Pada tahap iterasi ini, kami sekaligus menghitung jumlah dari jarak antar setiap cluster.
2. Simpan set of cluster ke dalam cluster history
   Pada tahap ini kami menyimpan data set of cluster yang kami iterasi pada tahap 1 sekaligus dengan jumlah dari jarak antar setiap clusternya. Cluster history ini yang

nantinya dapat digunakan saat user meminta set of cluster dengan jarak antar cluster < suatu threshold atau banyaknya cluster == suatu angka

3. Buat set of cluster baru
   Setelah terpilih dua cluster dengan jarak terdekat pada tahap 1. Kami membuat set of cluster baru dengan cara menggabungkan kedua cluster tersebut.
4. Ulangi
   Untuk menyelesaikan proses clusterisasi, ulangi tahap 1, 2, 3, dan 4 sampai jumlah cluster == 1.
5. Get result
   Proses pada tahap 1, 2, 3, dan 4 hanya akan menghasilkan cluster history. User harus meminta kemapa model untuk mendapat set of cluster (output) yang memenuhi suatu kondisi tertentu (i.e. jumlah dari jarak cluster < threshold atau banyaknya cluster == suatu nilai)

B. Hasil Eksekusi

Hasil percobaan yang dilakukan menggunakan algoritma yang telah dibuat terhadap dataset iris adalah sebagai berikut :

Parameter	Algoritma					
	k-means	Agglomerative				
		single	multiple	avg	avg-group	rata-rata
Percobaan 1						
Akurasi	0.89	0.67	0.89	0.90	0.69	0.79
FMI Iris-setosa	1.00	1.00	1.00	1.00	1.00	1.00
FMI Iris-versicolor	0.86	0.71	0.87	0.87	0.28	0.68
FMI Iris-virginica	0.83	0.14	0.82	0.84	0.72	0.63
Rerata Silhouette	0.55	0.36	0.55	0.55	0.48	0.49
Percobaan 2						
Akurasi	0.89	0.67	0.89	0.90	0.69	0.79
FMI Iris-setosa	1.00	1.00	1.00	1.00	1.00	1.00
FMI Iris-versicolor	0.85	0.71	0.87	0.87	0.28	0.68
FMI Iris-virginica	0.89	0.14	0.82	0.84	0.72	0.63
Rerata Silhouette	0.55	0.36	0.54	0.54	0.48	0.49
Percobaan 3						
Akurasi	0.52	0.67	0.89	0.90	0.69	0.79
FMI Iris-setosa	0.71	1.00	1.00	1.00	1.00	1.00
FMI Iris-versicolor	0.00	0.71	0.87	0.87	0.28	0.68
FMI Iris-virginica	0.72	0.14	0.82	0.84	0.72	0.63
Rerata Silhouette	0.48	0.36	0.54	0.54	0.48	0.49

C. Analisis

Dari tabel hasil eksekusi pada bagian B, kita dapat melihat bahwa secara keseluruhan menggunakan rata-rata, performa algoritma K-means lebih baik daripada algoritma Agglomerative. Namun harus kita pertimbangkan juga bahwa hasil dari algoritma K-Means dapat dikatakan kurang pasti karena state awal yang random. Selain itu, dapat kita lihat juga bahwa bila kita bandingkan dengan hasil terbaik Agglomerative (menggunakan fungsi jarak average) hasil Agglomerative lebih baik.

D. Pembagian Tugas

Nama	NIM	Pembagian Tugas
Muhammad Akmal	13517028	Implementasi K-Means Clustering, Laporan
Haris Salman Al- Ghifary	13517052	Visualisasi, analisis, Laporan
Hafidh Rendyanto	13517061	Implementasi Agglomerative Clustering, Laporan
Muhammad Nurdin Husen	13517112	Implementasi FMI Implementasi Silhouette Laporan