Nama : Edy Santoso
Kelas : IF 40-08
Nim : 1301164147

Tugas 2

1. **Kekurangan K-means :**
   a. metode k-means akan tidak optimal apabila akan mengklasterisasi data yang terdapat pencilan atau outlier. Karena apabaila terdapat pencilan, data ini akan mendistorsi nilai rerata cluster secara dramatis.
   b. angat sensitive pada pembangkitan titik pusat awal secara random
   c. sangat sulit mencapai global maksimum

   Kelebihan K-means :
   a. selalu konvergen atau mampu melakukan klasterisasi
   b. tidak membutuhkan operasi matematis yang rumit.
   c. beban komputasi yang relative lebih ringan sehingga klasterisasi lebih cepat.

   Contoh :

   Penggunaan metode klasterisasi dapat digunakan dalam berbagai hal dikehidupan nyata. Misalnya dalam dunia bisnis, metode ini dapat digunakan untuk menentukan customer segmentation. Misalnya dalam dunia provider telepon, metode ini dapat digunakan untuk mengelompokkan pelanggan sehingga dapat menawarkan paket promosi yang tepat. Proses identifikasi bias melalui history dari pembelian pulsa misalnya, atau pengambilan paket internet dan lain lain. Dengan menggunakan metode ini, dapat membantu perushaan untuk mempunyai target customer yang lebih spesifik.

2. ***Agglomerative Hierarchical Clustering***
   Metode *Agglomerative Hierarchical Clustering* adalah metode yang terdapat pada hierarchical clustering yang mempunyai konsep dengan meletakkan setiap objek yang ada sebagai sebuah cluster tersendiri. Kemudian setelah menjadikan setiap objek sebagai klaster, akan dilanjutkan dengan menggabungkan setiap klaster tadi menjadi klaster yang lebih besar lagi. Ada beberapa pendekatan yang dapat dilakukan dengan metode ini yaitu single link, multi link, group average dan mean.

   Contoh:

   Dalam kasus identifikasi hewan, apakah suatu hewan mempunyai DNA yang mirip dengan hewan yang lainnya. Misalnya apakah panda mempunyai kemiripan dengan beruang atau rakun. Dengan metode ini dapat diketahui apakah mereka adalah hewan yang sejenis dengan menganalisis dari DNA sequences disetiap hewan.

Laporan Klasterisasi Data dengan Menggunakan metode Self Organizing Map (SOM)

1. Deskripsi Masalah

Terdapat datset yang memiliki 600 sampel data dan dua atribut. Data yang tersedia mempunyai atribut kontinu. Dataset yang tersedia tidak memiliki label kelas disetiap sampel datanya. Oleh karena itu, diperlukan sebuah model sistem klasterisasi untuk mengelompokkna sampel data ek beberapa klaster yang optimum dengan menggunakan Self Organizing Map.

2. Metode Penyelesaian

Permasalahan yang sudah dideskripsikan diatas akan diselesaikan dengan menggunakan salah satu algoritma klastering yaitu Self Organizing maps. Metode ini akan berfokus pada jumlah neuron dan peletakannya dalam penentuan klaster. Jumlah neuron yang diambil akan menghasilkan jumlah klaster yang sama.

Langkah -langkah dalam metode Self organizing Maps sebagai berikut:

1. menentukan jumlah neuron yang akan dijadikan klaster nantinya. Dalam sistem yang dibuat akan dicoba beberapa neuron dari 6 neuron hingga 20.
2. setelah menentukan neuron, maka generate secara random atribut dari neuron. Generate angka untuk atribut pada neuron dibatasi antara 0-17. Hal ini dikarenakan rentang atribut pada objek di dataset mempunyai batas bawah 0 dan batas atas17.
3. setelah men-generate atribut untuk neuron, hitung jarak sampel data dengan setiap neuron yang telah di generate. Lakukan perhitungan dengan Euclidean distance.

$$S = \sqrt{\Sigma(xi - yi)^2}$$

3. kemudian. Bandingkan dan pilih sampel data yang mempubnyai jarak terdekat dengan neuron.
4. setelah itu, masukkan S yang telah dipunya pada step sebelumnya pada persamaan T.

$$T_{j,I(\mathrm{x})} = \exp\left(-S^2_{j,I(\mathrm{x})} / 2\sigma^2\right)$$

S adalah jarak pada step sebelumnya, lalu σ adalah sigma yang diinisiasi awal 2.sigma akan berubah pada setiap iterasi yang dilakukan.

5. setelah mendapatkan T, maka langkah selanjutnya dengan mencari Δw . Δw sendiri mempunyai persamaan seperti berikut:

$$\Delta w_{ji} = \eta(t) \cdot T_{j,I(\mathrm{x})}(t) \cdot \left(x_i - w_{ji}\right)$$

Dalam proses ini akan dihasilkan titik neuron yang baru. Yang mana xi adalah titik objek, kemudian w merupakan titik neuron. Dan n(t) merupakan learning rate yang diinisiasi 0.1 learning rate akan selalu update setiap iterasinya.

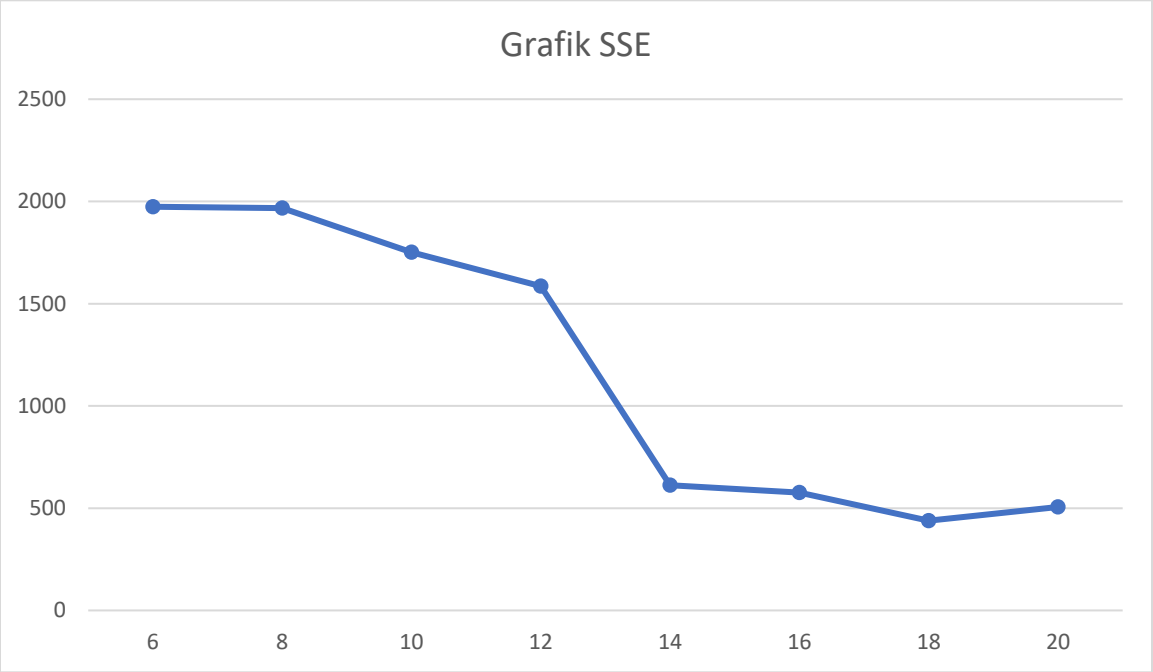
6. pada proses 6, proses satu iterasi telah dilakukan. Apabila akan dilakukan pengulangan hingga iterasi tertentu, maka sigma dan learning rate harus diubah dengan persasmaan berikut:

$$\sigma(t) = \sigma_0 \exp(-t/\tau_\sigma)$$

$$\eta(t) = \eta_0 \exp(-t/\tau_\eta)$$

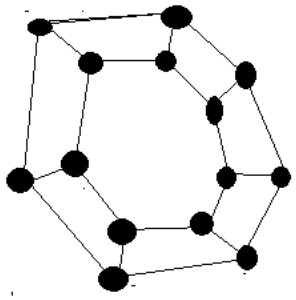
7. setelah proses iterasi selesai, maka akan menghasilkan titik titik neuron yang dianggap palig optimum. Maka lakukan pengecekan kedekatan setiap sampel data dengan neuron yang terdekat. Neuron yang terdekat dengan sampel data, akan menjadi klaster pada sampel data tersebut.
8. untuk memilih neuron yang maksimal dapat dilakukan dengan melakukan pengecekan pada SSE untuk setiap jumlah neuron yang diambil. Jumlah neuron yang diambil adalah yang mempunyai rentang SSE antar setiap neuron yang paling besar atau signifikan.

3. Hasil dan Pembahasan



Dapat dilihat pada grafik diatas, rentang SSE pada neuron 12 dan 14 adalah rentang yang paling signifikan disbanding yang lainnya. Berdasarkan hal itu, maka jumlah neuron 14 adalah yang paling optimum atau dengan kata lain jumlah kalaster yang optimum adalah 14.

Hasil klasterisasi juga dipengaruhi dari dimensi neutron yang dibangun. Dalam kasus ini, penulis menggunakan 3 dimensi dalam membangun 14 neutron untuk klasifikai. Adapaun gambaran letak neutron adalah sebagai berikut.



Neuron yang dibangun adalah neuron 3 dimensi, yang mana, setap neuron mempunyai 3 tetangga didekatnya. dimensi neuron sangat berpengaruh pada jumlah klaster optimum yang akan dihasilkan. Apabila dibangun pada dua dimensi, maka akan berbeda hasilnya, karena tetangga yang dipunya setiap neuron hanya ada dua.

Adapun bebrapa sampel hasil pemrosesan klasterisasi sebagai berikut :

1	9.802	10.132	10
2	10.35	9.768	6
3	10.098	9.988	6
4	9.73	9.91	10
5	9.754	10.43	10
6	9.836	9.902	10
7	10.238	9.866	6
8	9.53	9.862	10
9	10.154	9.82	6
10	9.336	10.456	10
11	9.378	10.21	10
12	9.712	10.264	10
13	9.638	10.208	10
14	9.518	9.956	10
15	10.236	9.91	6

Gambar diatas adalah sampel keluaran sistem yang dihasilkan, dimana data pada sampel pertama terdapat pada klaster 10. Untuk hasil lainnya, telah dilampirkan pada file excel dalam folder ini.