Tugas Besar machine Learning Self-Organizing-Map (SOM)

Badrus Shoolehk Al Ar Fanny

IF-40-04

Fakultas Informatika, Universitas Telkom
Jl. Telekomunikasi no.1 Terusan Buah Batu, Kab. Bandung, Jawa Barat
badrussholehaxel @gmail.com

No 1. K-Means clustering

Kelebihan dan kekurangan

Kelebihan	kekurangan
Menggunakan prinsip yang sederhana, dapat dijelaskan dalam non-statistik	Karena menggunakan k buah acak, tidak di jamin untuk menemukan kumpulan cluster yang optimal
Waktu yang dibutuhkan untuk menjalankan nya relatif cepat	dapat terjadinya curse of dimensionality, apabila jarak antara cluster yang satu dengan yang lain memiliki banyak dimesi.
Sangat fleksibel, dapat dengan mudah diadaptasi.	Tidak optimal digunakan untuk data yang jumlahnya terlalu banyak sampai bermiliyar.
Sangat umum digunakan	

Data	Attribut / fitur	
	X	Y
A	5,09	5,80
B	3,24	5,90
C	1,68	4,90
D	1,00	3,17
E	1,48	1,38
F	2,91	0,20
G	4,76	0,10
H	6,32	1,10
I	7,00	2,83
J	6,52	4,62

Contoh penerapan K-Means Cluster

Akan dilakukan pemartisian data terhadap data diatas sebanyak 2 partisi, maka tahapannya dalah sebagai berikut :

K = 2, (K = Jumlah Cluster)
Nilai awal centroid (1.48,1.38) untuk partisi 0, dan centroid (4.76,0.10) untuk partisi 1

$$D(p,c)_n = \sqrt[2]{\sum_{i=0}^{n}(p_i - c_i)^2}$$

Maka perhitungan Euclidean Distance adalah dimana ***p*** adalah data, ***c*** adalah

centroid, n adalah jumlah data, i adalah iterasi.

Hasil perhitungan jarak minimum A adalah

 Hasil perhitungan jarak minimum B adalah

$$D(p,c)_0 = \sqrt{(3.24-1.48)^2 + (5.90-1.38)^2} \approx 4.851$$

$$D(p,c)_1 = \sqrt{(3.24-4.76)^2 + (5.90-0.10)^2} \approx 5.996$$

 Hasil perhitungan jarak keseluruhan

Med	Cluster 0	Cluster 1
A	1	0
B	1	0
C	1	0
D	1	0
E	1	0
F	0	1
G	0	1
H	0	1
I	0	1
J	0	1

 Hasil perhitungan partisi diambil dari jarak minimum, sehingga didapat sebagai berikut : Contoh pada A, X = 5.707 lebih kecil dari Y = 5.710, maka A termasuk ke dalam Cluster 0. Begitu juga dengan F, X = 1.854 lebih besar dari Y = 1.853, maka B masuk ke dalam Cluster 1.

 Setelah data di partisi, maka selanjutnya nilai centroid harus dihitung ulang untuk menentukan jarak minimum yang baru, berikut perhitungan centroid baru :

$$c_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{p_j \in S_i^{(t)}} p_j$$

$$c_0 = \left(\frac{(5.09+3.24+1.68+1.00+1.48)}{5}, \frac{(5.80+5.90+4,90+3.17+1.38)}{5}\right) \approx (2.498, 4.23)$$

$$c_1 = \left(\frac{(2.91+4.76+6.32+7+6.52)}{5}, \frac{(0.2+0.1+1.1+2.83+4.62)}{5}\right) \approx (5.502, 1.77)$$

Hitung jarak minimumnya kembali dengan menggunakan centroid yang baru, sehingga di dapat hasilnya sebagai berikut.

Klasifikasikan kembali data berdasar jarak minimum diatas.

Karena tidak ada data yang berpindah ke cluster yang berbeda, sehingga iterasi kita cukupkan sampai dengan nilai centroid akhir : **(2.50,4.23), (5.50,1.77)**

$$D(p,c)_0 = \sqrt{(5.09-1.48)^2 + (5.80-1.38)^2} \approx 5.707$$

$$D(p,c)_1 = \sqrt{(5.09-4.76)^2 + (5.80-0.10)^2} \approx 5.710$$

No 2. Hierarchical Clustering

adalah metode analisis kelompok yang berusaha untuk membangun sebuah hirarki kelompok data.

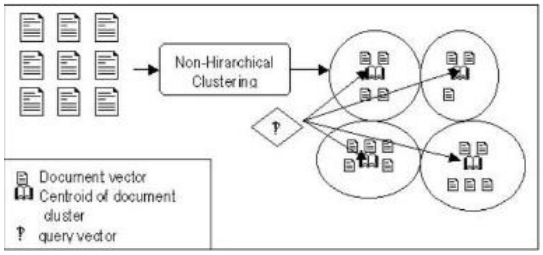
Strategi pengelompokannya umumnya ada 2 jenis yaitu **Agglomerative (Bottom-Up)** dan **Devisive (Top-Down)**.

Contoh kasus

Clustering HierarchicalMetode pembentukan cluster biasanya dikategorikan menurut tipe dari struktur clusteryang dihasilkan. Secara umum metode clusterterbagi menjadi dua, yaitu metode Non-ierarchical Clustering(klasteringnon-hirarkhis) dan metode Hierarchica lClustering(klastering hirarkhis).Metode non-hirarkhis disebut juga metode partisi, yaitu membagi serangkaian data yang terdiri dari n obyek ke dalam k cluster(k<n) yang tidak saling tumpang-tindih (overlap), dimana nilai k telah ditentukan sebelumnya.

Salah satu prosedur pengelompokkan pada non-hirarkhis adalah dengan menggunakan metode k-means. Metode ini merupakan metode pengelompokkan yang bertujuan untuk mengelompokkan objek sedemikian hingga jarak tiap-tiap objek kepusat kelompok didalam suatu kelompok adalah minimum.Pembentukan clusterdokumen dalam Sistem Temu Kembali Informasi dengan metode non-hirarkhis adalah sebagi

berikut Membandingkan ciri-ciri identifikasi (identifier) suatu dokumen dengan dokumen lain yang ada dalam koleksi dan mengelompokkan dokumen-dokumen yang memiliki serangkaian ciri-ciri identifikasi yang serupa ke dalam satu cluster.b.Pada setiap clusterdokumen yang dihasilkan, dipilih sebuah unsur yang dapat mewakili seluruh dokumen yang ada dalam clusteryang bersangkutan yang disebut centroid. Centroid atau perwakilan clusteradalah sebuah record yang dapat mewakili ciri-ciri atau karakteristik dokumen dalam sebuah cluster.c.Proses penelusuran dilakukan dalamdua tahap, yaitu: 1) membandingkan query dengan centroid pada masing-masing clusterdokumen; 2) mencocokkan query dengan masing-masing dokumen dalam clusteryang mengandung centroid yang paling sesuai.



Proses pembentukan clusterdokumen dan penelusuran tersebut dapat diilustrasikan seperti pada Gambar 1 di bawah ini :Gambar 1 ClusterDokumen dengan MetodeNon-hirarkhisMetode clusteryang kedua adalah metode Hierarchical Clustering(klastering hirarkhis). Metode pengelompokkan hirarkhis biasanya digunakan apabila belum ada informasi jumlah kelompok yang akan dipilih. Arah pengelompokkan bisa bersifat divisive (top to down) artinya dari 1 clustersampai menjadi k buah clusteratau bersifat agglomerative (bottom up) artinya dari n cluster(dari n-buah data yang ada) menjadi k buah cluster. Teknik hirarkhis (hierarchical methods) adalah teknik clustering membentuk kontruksi hirarki atau berdasarkan tingkatan tertentu seperti struktur pohon. Dengan demikian proses pengelompokkannya dilakukan secara bertingkat atau bertahap.Hierarchical Clussteringadalah salah satu algoritma clustering yang dapat digunakan untuk meng-clusterdokumen (document clustering). Dari teknik Hierarchical Clustering, dapat dihasilkan suatu kumpulan partisi yang berurutan, dimana dalam kumpulan tersebut terdapat:a.Cluster-clusteryang mempunyai poin poin individu. Cluster-clusterini berada di level yang paling bawah.b.Sebuah clusteryang didalamnya terdapat poin poin yang dipunyai semua clusterdidalamnya. Single clusterini berada di levelyang paling atas.Pembentukan clusterdokumen dalam Sistem Temu Kembali Informasi dengan metode hirarkhis adalah sebagi berikut:a.Mengidentifikasi dua dokumen yang paling miripdan menggabungkannya menjadi sebuah cluster.b.Mengidentifikasi dan menggabungkan dua dokumen yang paling mirip berikutnya menjadi sebuah clustersampai semua dokumen tergabung dalam cluster-clusteryang terbentuk.b.Proses penelusuran dokumen dilakukan dengan cara mencocokkan query dengan centroid. Centroid merupakan dokumen parentpada masing-masing clusterdokumen. Berikutnya dokumen yang berada dalam satu clusterdengan centroidakan ditampilkan sebagai hasil query.

No 3. Self Organizing Map

Problem :

Diberikan sebuah dataset berisi600 objek data yang memiliki dua atribut tanpalabel kelas. Bangunlah sebuah model klasterisasi (clustering) menggunakan metode

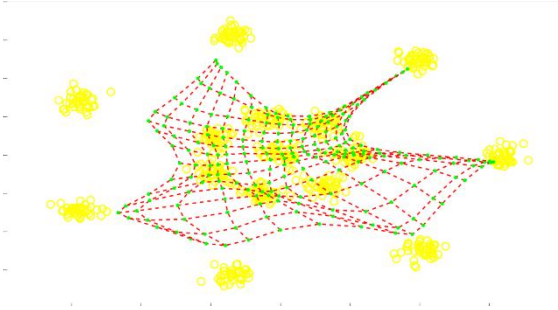
SelfOrganizing Map (SOM) untuk menghasilkan sejumlah klaster yang paling optimum. Strategi Penyelesaian Masalah : Strategi yang digunakan dengan mengimplementasikan metode Self OrganizingMap kedalam program sehingga program tersebut dapat menentukan label dari 600 datayang diberikan. Dalam program yang saya buat, saya membangkitkan neuron sebanyak 8 neuron dengan weight dari masing-masing neuron di random dari -15 sampai 15 di awalpenentuan neuronnya, setelah itu setiap data akan di hitung jaraknya ke masing-masingneuron dan neuron yang terdekat dari data tersebut akan di pilih menjadi neuron pemenang,Stelah itu akan di hitung jarak antar neuronnya ke neuron pemenang jika jaraknya lebihkecil dari 10 maka neuron tersebut akan di pilih menjadi neuron tetangganya, untukmenghuitng jarak menggunakan Euclidian Distance, Setelah mendapatkan jarak dari antarneuron maka cari Tnnya dengan rumus EXP(-(Sn^2)/2*sigma), setelah itu akan dicariwnnya dengan rumus Lr*Tn*(x-n), dimana Lr dan sigma adan selalu berubah di setiapiterasinya dan terakhir maka mengubah nilai weight dari neuronnya, setelah itumengkelompokan data ke neuron terdekatnya.

Analisis :

Berdasarkan penjelasan yang saya jabarkan maka dapat diketahui bahwa pembangkitkan berapa banyak neuron sangat penting untuk mendapatkan klaster yangoptimum, dan pengambilan tetangga dari neuron pemenangnya juga sangat berperanpenting.

Hasil Percobaan :

Ini adalah hasil Percobaan yang telah malalui SOM (*Self Orgnaizing Map*)



Ini adalah hasil percobaan yang sebelum terjadinya SOM