PENGELOMPOKAN DATA DENGAN METODE KLASTERISASI HIRARKI

Oleh
Luh Joni Erawati Dewi
Jurusan Manajemen Informatika, FTK, Undiksha

Abstrak

Pengelompokan data sangat diperlukan untuk mengetahui karakteristik kemiripan data. Tetapi, seringkali ditemui kesulitan untuk mengelompokkan data yang tidak berlabel. Untuk menyelesaikan persoalan ini bisa digunakan metode klasterisasi hirarki terhadap data yang ada. Sebuah objek dengan objek lainnya akan berada pada kelompok yang sama jika mempunyai kemiripan yang dekat dibandingkan dengan objek yang ada pada kelompok lainnya. Pada tulisan ini akan dibahas metode pengelompokan data dengan menggunakan algoritma klasterisasi hirarki.

Kata-kata kunci : algoritma klasterisasi hirarki, kemiripan data, pengelompokan data

Abstract

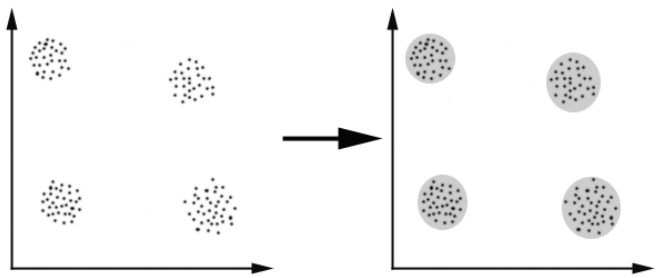
Grouping of data is needed to determine the characteristics of data similarity. But, we often found it difficult to classify data that are not labeled. To resolve this issue could use a hierarchical clustering method to the data. An object with other objects will be in the same group if it has a close similarity compared with the existing objects in other groups. This paper will discuss the method of grouping data using a hierarchical clustering algorithm.

Key words: hierarchical clustering algorithm, similarity data, grouping the data

I. PENDAHULUAN

Pengelompokan data bisa dilakukan dengan berbagai cara. Untuk mengelompokkan data yang tidak berlabel bisa diselesaikan dengan metode klasterisasi. Masalah klasterisasi berkaitan dengan penemuan struktur pada kumpulan data yang tidak berlabel. Definisi klasterisasi secara sederhana adalah proses mengorganisasi objek menjadi kelompok-kelompok yang anggotanya sama dalam beberapa hal. Jadi klaster adalah

kumpulan objek yang mempunyai kesamaan tertentu pada suatu klaster dan mempunyai perbedaan dengan objek pada klaster lainnya. Contoh sederhananya dapat dilihat pada Gambar 1.



(a) (b)

Gambar 1. Contoh klasterisasi

Pada Gambar 1a, dengan mudah dapat diidentifikasikan empat buah klaster ke mana data dapat dibagi(Gambar 1b). Kriteria persamaan di sini adalah jarak. Dua objek atau lebih terdapat pada klaster yang sama jika objek-objek tersebut berdekatan menurut ukuran jarak (pada kasus ini jarak geometri). Hal ini disebut dengan klasterisasi berdasar jarak. Jenis klaster yang lainnya adalah klasterisasi konsep, yaitu: dua objek atau lebih terdapat pada klaster yang sama jika objek-objek tersebut mempunyai konsep yang sama. Dengan kata lain, objek dikelompokkan menurut kesesuaiannya pada konsep deskriptif, tidak menurut ukuran persamaan sederhana.

Tujuan klasterisasi adalah menentukan pengelompokan intrinsik pada data yang tidak berlabel. Untuk memutuskan apakah sebuah klaster sudah bagus atau tidak, bukanlah keputusan yang mudah. Tidak ada kriteria yang pasti, yang bebas dari tujuan akhir klasterisasi. Konsekuensinya pengguna harus menyediakan kriteria ini sehingga hasil klasterisasinya akan memenuhi kebutuhan pengguna.

Sebagai contoh, mungkin pengguna tertarik pada penemuan kelompok yang homogen(yang disebut juga reduksi data), atau dalam hal penemuan klaster alami(kelompok alami) dan mendeskripsikan propertis yang dimilikinya, atau pada penemuan kelompok yang bermanfaat dan

sesuai (kelas data ”*useful*”) atau pada penemuan data yang tidak biasa(deteksi anomali/*outlier*).

Makalah ini menyajikan ringkasan mengenai klasterisasi hirarki, algoritma hirarki aglomeratif, dan sebuah contoh kasus pengelompokan data dengan menerapkan algoritma hirarki aglomeratif. Ringkasan ini diharapkan bisa bermanfaat untuk memperkaya pengetahuan khususnya metode pengelompokkan data.

II. PEMBAHASAN

2.1 Klasterisasi Hirarki

Klasterisasi hampir mirip dengan klasifikasi dalam hal pengelompokan data. Bedanya adalah klasterisasi tidak menentukan kategori kelompok sebelumnya(Dunham, 2003).

Metode klasterisasi hirarki ini bekerja dengan mengelompokkan data yang terdekat berdasarkan kriteria tertentu menjadi satu klaster. Hal ini dilakukan secara berulang-ulang sampai pada akhirnya terbentuk sebuah klaster tunggal dengan jumlah item sebanyak jumlah item data yang diberikan. Metode ini disebut klasterisasi hirarki aglomeratif. Hirarki aglomeratif sangat sesuai digunakan untuk aplikasi seperti toksonomi yang secara alami mengandung struktur hirarki. Selain itu, terdapat beberapa kajian yang menuliskan bahwa algoritma hirarki aglomeratif dapat menghasilkan kualitas klaster yang lebih baik(Tan dkk, 2006).

2.2 Algoritma Klasterisasi Hirarki

Berikut ini adalah cara kerja algoritma hirarki aglomeratif. Diberikan sekumpulan N item yang akan diklaster, dan sebuah matrik N*N yang menyatakan jarak antar item pada N, proses utama dari klasterisasi hirarki adalah (sesuai dengan S.C Johnson, 1967 dalam Moore,2007)

1. Mulai dengan membuat klaster sebanyak N, masing-masing klaster mempunyai sebuah item. Misalnya jarak antar klaster sama dengan jarak antar item yang dikandungnya.
2. Cari sepasang klaster yang jaraknya terdekat, dan jadikan sebuah klaster baru. Jadi sekarang kita mempunyai N-1 klaster.
3. Hitung jarak antara klaster yang baru dengan masing-masing klaster yang lainnya
4. Ulangi langkah 2 dan 3 sampai semua item menjadi sebuah klaster dengan N item. Tentunya tidak ada gunanya mempunyai N item yang dikelompokkan menjadi satu klaster besar.

Langkah 3 dapat dilakukan dengan tiga(3) cara yaitu *single-linkage*, *complete-linkage*, dan *average-linkage*. Pada klasterisasi *single-linkage* (atau metode minimum), jarak antara satu klaster dengan klaster lainnya sama dengan jarak terdekat antara sebuah item yang terdapat dalam satu klaster dengan item dari klaster lainnya. Pada klasterisasi *complete-linkage* (disebut juga metode maksimum) jarak antara satu klaster dengan klaster lainnya sama dengan jarak terjauh antara sebuah item yang terdapat dalam satu klaster dengan item dari klaster lainnya. Pada klasterisasi *average-linkage*, jarak antara satu klaster dengan klaster lainnya sama dengan jarak rata-rata antara sebuah item yang terdapat dalam satu klaster dengan item dari klaster lainnya(Tan,dkk, 2006).

Klasterisasi hirarki disebut aglomeratif karena menggabungkan dua klaster secara iteratif. Berikutnya yang dibahas adalah klasterisasi *single-linkage*.

Algoritma klasterisasi *Single-Linkage* mengandung skema aglomeratif yang menghapus baris dan kolom pada matrik proksimiti. Baris dan kolom yang dihapus ini merupakan klaster yang digabung menjadi klaster yang baru.

Matrik proksimiti N*N adalah D=[d(i,j)]. Klasterisasi ditandai dengan angka berurut 0,1,...,(n-1). L(k) adalah level klasterisasi ke k. Klaster dengan urutan angka m dinyatakan dengan (m) dan proksimiti antara

klaster r dan s dinyatakan dengan d[(r), (s)]. Algoritma *single-linkage* disusun dari langkah-langkah berikut.

1. Mulai dengan klasterisasi yang mempunyai level L(0)=0 dan m=0
2. Temukan pasangan klaster dengan kesamaan terkecil pada klaster sekarang, misalkan (r) dan (s), sesuai
   D[(r),(s)]=min d[(i),(j)] di mana nilai minimum diperoleh dari semua pasangan klaster.
3. Naikkan m=m+1. Gabungkan klaster (r) dan (s) menjadi klaster tunggal untuk membentuk klaster m. Set level dari klaster ini menjadi L(m) =d[(r),(s)]
4. *Update* matrik proksimiti, D, dengan menghapus baris dan kolom yang berkorespondensi pada klaster (r) dan (s) dan tambahkan sebuah baris dan kolom yang berkorespondensi pada klaster baru. Proksimiti antara klaster baru, dinyatakan (r,s) dan klaster lama (k) didefinisikan dengan:
   d[(k), (r,s)] = min d[(k),(r)], d[(k),(s)]
5. Jika semua objek berada pada satu klaster yang sama, berhenti. Sebaliknya, kembali ke langkah 2.

2.3 Contoh Kasus

Sebagai contoh dapat dilihat pada kasus berikut ini. Akan dibuat pengelompokan data hasil tes lari mahasiswa dengan metode klasterisasi hirarki. Kriteria jarak yang digunakan di sini adalah beda waktu yang diperlukan oleh masing-masing mahasiswa untuk menempuh enam kali putaran keliling lapangan dalam hitungan detik. Metode ini menggunakan *single-linkage* sesuai dengan langkah-langkah di atas.

Input matrik jarak(L=0 untuk semua klaster). Matrik yang terbentuk dapat dilihat pada Tabel 1.

Tabel 1. Matrik jarak hasil tes mahasiswa

	M1	M2	M3	M4	M5	M6
M1	0	5	8	10	15	22
M2	5	0	10	9	13	25
M3	8	10	0	9	11	22
M4	10	9	9	0	6	26
M5	15	13	11	6	0	28
M6	22	25	22	26	28	0

Pasangan jarak terdekat adalah M1 dan M2, pada jarak=5. Keduanya digabung menjadi klaster tunggal disebut M1/M2. Level dari klaster baru adalah L(M1/M2) = 5 dan m=1. Kemudian dihitung jarak dari objek gabungan ini ke semua objek lainnya. Pada klasterisasi *single link*, aturannya adalah jarak antara objek gabungan ke objek lainnya sama dengan jarak terpendek dari suatu anggota pada klaster ke yang lainnya di luar objek. Jadi jarak dari M1/M2 ke M3 dipilih 8, yang merupakan jarak dari M1 ke M3, begitu seterusnya. Setelah muncul klaster M1/M2 diperoleh matrik seperti pada Tabel 2.

Tabel 2. Matrik setelah Penggabungan M1 dan M2

	M1/M2	M3	M4	M5	M6
M1/M2	0	8	9	13	22
M3	8	0	9	11	22
M4	9	9	0	6	26
M5	13	11	6	0	28
M6	22	22	26	28	0

Jarak hasil tes yang terdekat berikutnya adalah M4 dan M5, jadi M4 dan M5 digabung menjadi sebuah klaster baru. Level dari klaster baru adalah L(M4/M5) = 6 dan m=2.

Matrik yang terbentuk seperti terlihat pada tabel 3 berikut ini.

Tabel 3. Matrik setelah Penggabungan M4 dan M5

	M1/M2	M3	M4/M5	M6
M1/M2	0	8	9	22
M3	8	0	9	22
M4/M5	9	9	0	26
M6	22	22	26	0

Dari matrik di atas diperoleh jarak M1/M2 dengan M3 paling dekat, sehingga M1/M2 dan M3 menjadi sebuah klaster baru. Level dari klaster baru adalah L(M1/M2/M3) = 8 dan m=3.

Jarak hasil tes antar mahasiswa dihitung lagi, sehingga diperoleh matrik seperti terlihat pada Tabel 4. berikut ini.

Tabel 4. Matrik setelah Penggabungan M1/M2 dengan M3

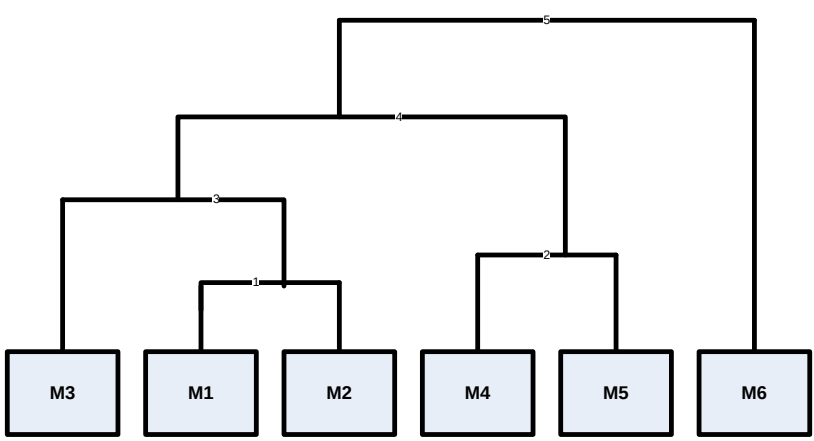
	M1/M2/M3	M4/M5	M6
M1/M2/M3	0	9	22
M4/M5	9	0	26
M6	22	26	0

Selanjutnya M1/M2/M3 dikelompokkan menjadi satu klaster dengan M4/M5. Level dari klaster baru = 9, dan m = 4. Matrik diperbarui lagi berdasarkan jarak hasil tes terbaru antar mahasiswa. Matrik yang terbentuk seperti terlihat pada Tabel 5.

Tabel 5. Matrik setelah Penggabungan M1/M2/M3 dengan M4/M5

	M1/M2/M3/M4/M5	M6
M1/M2/M3/M4/M5	0	22
M6	22	0

Akhirnya, gabungkan dua klaster terakhir pada level 22. Proses dapat disimpulkan dengan pohon hirarki seperti pada Gambar 2 berikut ini.



Gambar 2. Pohon hirarki data tes mahasiswa

Jika misalnya menginginkan data hasil tes mahasiswa dibagi menjadi tiga buah klaster maka klaster yang dihasilkan adalah:

Klaster 1: M1, M2, M3

Klaster 2: M4,M5

Klaster 3: M6

III. PENUTUP

Dari pembahasan di atas dapat disimpulkan bahwa klasterisasi hirarki aglomeratif dapat mengelompokkan data yang tidak mempunyai label. Namun, algoritma klasterisasi hirarki aglomeratif ini mempunyai kelemahan utama yaitu:

1. tidak efisien, pada setiap iterasi dilakukan perhitungan kriteria(jarak) antar objek dengan setiap anggota klaster yang ada sehingga kompleksitas waktunya setidaknya $O(n^2)$, di mana n adalah jumlah total objek.
2. apa yang sudah dikerjakan sebelumnya tidak bisa dibalik lagi.

DAFTAR PUSTAKA

Dunham, M.(2003). "*Clustering*", Introductory and Advanced Topics. USA: Pearson Education.

Moore, A: "K-means and Hierarchical *Clustering* - Tutorial Slides" http://www-2.cs.cmu.edu/~awm/tutorials/ kmeans.html, diakses tgl 22 April 2007

Tan, Pang-Ning, Steinbach Michael, Kumar, Vipin, (2006). "*Agglomerative Hierarchical Clustering*",*Introduction to Data Mining*. USA: Addison Weslay.