seri tulisan data mining

.: Analisis Gerombol - Bagian 1 :.

Konsep Dasar

Bagus Sartono

bagusco@gmail.com

May 15, 2016

Abstract

Pada seri tulisan ini akan dipaparkan beberapa hal dasar dan umum mengenai analisis gerombol (*cluster analysis*) dengan menitikberatkan pada pemahaman konsep dasar dan penjelasan proses algoritmanya. Beberapa aplikasi akan diberikan dengan uraian yang lebih terbatas. Kritik, saran dan pertanyaan terhadap materi tulisan ini dapat disampaikan melalui email pada alamat bagusco@gmail.com. Penulis akan sangat mengapresiasi berbagai masukan tersebut. Akhirnya, selamat membaca.

1 Pengantar

Adalah sifat dari manusia secara umum bahwa jika berhadapan dengan data yang semakin besar ukurannya, baik dari sisi banyaknya amatan maupun banyaknya variabel, maka akan semakin sulit menemukan pola yang terkandung pada data tersebut. Pola yang terkandung dalam data dapat berupa berbagai hal, yang salah satunya adalah adanya amatan-amatan yang bersifat mirip satu sama lain dilihat dari berbagai variabel yang tersedia dalam gugus data tersebut. Keberadaan amatan-amatan yang mirip ini dapat dinyatakan dalam kalimat lain sebagai keberadaan beberapa kelompok amatan, dimana antar amatan dalam satu kelompok tampak serupa satu dengan yang lainnya. Analisis gerombol (diterjemahkan dari *cluster analysis*) merupakan analisis yang berguna dalam mengidentifikasi keberadaan kelompok dalam gugus data yang berisi banyak amatan dan setiap amatan dicirikan oleh banyak variabel.

Analisis gerombol ini bermanfaat dalam banyak bidang terapan. Beberapa ilustrasi yang dapat disebutkan antara lain adalah

- Segmentasi Pelanggan Operator Telekomunikasi Seluler

  Sebuah perusahaan operator seluler di suatu negara dapat memiliki pelanggan individu

yang banyaknya mencapai ratusan ribu hingga jutaan orang. Menarik untuk melihat bagaimana pola penggunaan berbagai feature/produk yang ditawarkan oleh perusahaan tersebut. Andaikan untuk setiap pelanggan bisa diidentifikasi bagaimana penggunaan setiap produk misalnya durasi melakukan panggilan, durasi menerima panggilan, frekuensi mengirim pesan singkat, frekuensi menerima pesan singkat, penggunaan paket data internet, dan lain sebagainya. Dari sekian banyak pelanggan mungkin saja hanya ada 5 hingga 8 kelompok pelanggan saja berdasarkan variabel-variabel penggunaan tadi. Jika kelompok-kelompok tersebut telah teridentifikasi maka pihak operator dapat memanfaatkan informasi itu untuk menawarkan paket bundle produk yang lebih tepat.

- Segmentasi Nasabah Pengguna Debit Card

  Debit card telah menjadi salah satu kebutuhan penduduk di daerah urban di Indonesia, karena dengan ini banyak orang yang terbantu dalam melakukan pembayaran dibandingkan dengan bentuk uang tunai. Saat ini, terutama di perkotaan, telah banyak merchant (toko, outlet, restoran, supermarket, departemen store, rumah sakit, pom bensin dsb) yang memungkinkan untuk menerima pembayaran menggunakan kartu debit. Adalah menarik untuk melihat di tipe merchant seperti apa biasanya para nasabah pemegang kartu menggunakan kartu mereka. Sebagian barangkali hanya menggunakan di supermarket untuk belanja kebutuhan rumah tangga sehari-hari, namun ada pula yang menggunakannya lebih sering di restoran. Mengetahui kelompok nasabah berdasarkan tipe merchant ini penting bagi perusahaan sehingga dapat menawarkan atau menginformasikan program promo dengan lebih baik.

- Pengelompokan Desa/Kelurahan berdasarkan Keberadaan Prasarana

  Perencanaan atau penyusunan program tidak hanya miliki perusahaan untuk mengembangkan bisnis, namun juga dilakukan oleh pemerintah dalam merencanakan program pembangunan. Salah satu yang dikerjakan adalah pengembangan kawasan pedesaan di berbagai pelosok negeri ini. Penting untuk mengetahui berapa banyak grup desa di Indonesia. Yang dapat dikerjakan adalah mengumpulkan informasi mengenai berbagai karakteristik desa, misalnya keberadaan prasarana pendidikan, kesehatan, transportasi, dan ekonomi, dan selanjutnya melakukan pengelompokan desa sehingga dapat diketahui bagaimana profil umum dari desa-desa yang ada. Dengan informasi itu dapat ditentukan program pengembangan prasarana dengan lebih terfokus.

Konsep dasar pembentukan gerombol dalam analisis gerombol secara sederhana dapat dinyatakan sebagai berikut. Andaikan ada banyak objek dan setiap objek diamati pada beberapa variabel. Yang dilakukan oleh analisis gerombol adalah membentuk sejumlah kecil gerombol

dimana objek-objek gerombol bersifat lebih mirip dibandingkan objek-objek yang berada di gerombol yang lain. Karena itu, penting kemudian untuk membicarakan bagaimana kita dapat menilai tingkat kemiripan (*similarity* atau *proximity*) antara dua buah objek. Tingkat kemiripan antar objek ini selanjutnya dinilai menggunakan suatu ukuran yang disebut sebagai jarak (*distance*).

2 Jarak

Istilah jarak sering kita gunakan dalam diskusi yang terkait dengan lokasi. Dua lokasi yang berdekatan kita nyatakan dengan memberikan jarak yang bernilai kecil, sedangkan dua lokasi yang berjauhan kita nyatakan dengan nilai jarak yang besar. Misalnya, Bandung dan Bogor memiliki jarak 126 km sedangkan jarak antara Bandung dan Yogyakarta adalah 428 km. Dengan informasi tersebut kita dapat menyimpulkan bahwa dibandingkan ke Bogor, lokasi Yogyakarta lebih jauh kalau dipandang dari Bandung.

Bagaimana menghitung jarak antara dua kota? Pendekatan yang dilakukan adalah menggunakan sistem koordinat lintang (*lattitude*) dan bujur (*longitude*) dimana setiap titik di permukaan bumi dicirikan oleh masing-masing sebuah nilai untuk keduanya. Andaikan suatu titik A memiliki koordinat (x_1, y_1) dan titik kedua yaitu B memiliki koordinat (x_2, y_2) maka jarak antara keduanya antara lain dapat diwakili oleh panjang garis terpendek yang menghubungkan kedua titik, yang disebut sebagai Jarak *Euclid*.

Pada lokasi dalam ruang berdimensi dua seperti ilustrasi di atas, jarak Euclid dari A dan B diukur menggunakan

$$d_{AB} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{1}$$

Konsep serupa digunakan pada penentuan kedekatan atau kemiripan dua buah objek. Ingat bahwa dari setiap objek dapat dilakukan pengukuran atau pencatatan beberapa variabel. Misalnya dari suatu objek individu orang, dapat dikumpulkan nilai tinggi badannya, berat badanya, usianya, pendapatan per bulan-nya, dan lain-lain. Setiap variabel tersebut selanjutnya dapat berfungsi seperti halnya latitude dan longitude, namun dalam dimensi yang lebih banyak dan bukan merupakan posisi lokasi dalam pengertian ruang yang biasa kita kenal.

Andaikan setiap amatan diamati menggunakan p buah variabel yaitu $X_1, X_2, \ldots, X_p$ sehingga misalkan pengamatan atau objek A memiliki nilai dari variabel-variabel tersebut sebesar $(x_{1A}, x_{2A}, \ldots, x_{pA})$ serta objek pengamatan B memiliki nilai sebesar $(x_{1B}, x_{2B}, \ldots, x_{pB})$. Jarak Euclid dari objek A dan B kemudian dapat dituliskan sebagai

$$d_{AB} = \sqrt{(x_{1A} - x_{1B})^2 + (x_{2A} - x_{2B})^2 + \ldots + (x_{pA} - x_{pB})^2} \tag{2}$$

atau

$$d_{AB} = \sqrt{\sum_{i=1}^{p} (x_{iA} - x_{iB})^2} \tag{3}$$

atau bisa juga dituliskan dalam bentuk notasi vektor sebagai

$$d_{AB} = \sqrt{(a-b)^T (a-b)} \tag{4}$$

dengan a dan b masing-masing adalah vektor berikut: $a = (x_{1A}, x_{2A}, \ldots, x_{pA})^T$ dan $b = (x_{1B}, x_{2B}, \ldots, x_{pB})^T$.

Terdapat beberapa formula lain penghitungan jarak, namun sementara tidak kita bahas disini. Pada bagian-bagian berikutnya akan disinggung kembali formula-formula tersebut.

Dengan penjelasan di atas, sudah dapat dibayangkan bahwa jika kita memiliki banyak amatan/objek dan pada setiap objek diamati beberapa variable maka kita dapat menghitung kedekatan antar amatan. Kedekatan ini yang menjadi dasar utama proses pembentukan kelompok-kelompok objek pada analisis gerombol.

3 Jenis-Jenis Analisis Gerombol

Terdapat beberapa pembagian jenis-jenis analisis gerombol. Meskipun tidak semuanya akan dibahas pada seri tulisan ini, namun kami menilai ada baiknya memiliki pengetahuan mengenai pembagian tersebut.

Pembagian yang pertama adalah pembagian berdasarkan karakteristik keanggotaan objek dalam suatu gerombol. Pada pembagian ini, analisis gerombol dapat terbagi menjadi *fuzzy* dan *non-fuzzy*. Penggerombolan yang bersifat non fuzzy memiliki pengertian bahwa keanggotaan suatu objek ke dalam suatu kelompok bersifat tegas: ya atau tidak. Sedangkan pada analisis yang bersifat fuzzy, keanggotaan dinyatakan dalam bentuk berapa peluang setiap objek dikategorik dalam suatu gerombol dan total dari peluang tersebut adalah 1.

Pembagian yang kedua adalah berdasarkan ekskusifitas keanggotaan. Dalam pembagian ini, analisis gerombol dikategorikan bersifat eksklusif jika suatu objek hanya dikategorikan ke dalam satu dan hanya satu gerombol. Sebaliknya, jika satu objek dapat masuk ke dalam lebih dari satu gerombol maka disebut analisis gerombol yang non-ekslusif.

Yang berikutnya adalah pembagian analisis gerombol berdasarkan proses pembentukan gerombolnya. Ini adalah merupakan pembagian yang paling banyak disebutkan dalam berbagai literatur yang mengulas konsep dasar analisis gerombol. Pada pembagian ini, analisis gerombol dibagi menjadi analisis gerombol yang pembentukan gerombolnya bersifat berhirarki (*hierarchical clustering*) dan yang bersifat tidak berhirarki (*non-hierarchical clustering*). Analisis gerombol berhirarki dilakukan dengan pada awalnya menganggap setiap objek sebagai satu

buah gerombol yang beranggotakan satu objek, dan selanjutnya menggabungkan dua objek yang berjarak paling dekat menjadi sebuah gerombol yang lebih besar. Demikian seterusnya hingga terbentuk satu buah gerombol besar yang memuat semua objek asal di dalamnya. Sedangkan analisis gerombol tak berhirarki bekerja secara langsung memisahkan banyak objek asal menjadi sejumlah kecil gerombol, sehingga ada juga yang menyebutkan sebagai analisis gerombol yang bersifat *partitional*.

4 Penutup

Pada bagian-bagian selanjutnya dari seri tulisan ini akan difokuskan pada pemaparan analisis gerombol berhirarki dan tak berhirarki menggunakan konsep jarak, dengan karakteristik penggerombolan yang bersifat non-fuzzy dan eksklusif.