Klasterisasi Data Kategorikal dengan Menggunakan Algoritma Modes Linkage

Tita Karlita
Program Studi Teknik Informatika
Politeknik Elektronika Negeri Surabaya
Kampus ITS Sukolilo Surabaya 60111
tita@eepis-its.edu

Abstrak

Klasterisasi adalah salah satu operasi penting dalam bidang data mining. Salah satu algoritma klasterisasi hirarki untuk operasi klasterisasi adalah centroid linkage. Namun penggunaan algoritma ini hanya terbatas pada data numerik. Dalam penelitian ini dirancang suatu algoritma pengembangan dari algoritma centroid linkage yang bekerja pada data kategorikal yaitu algoritma modes linkage. Ujicoba dilakukan dengan menggunakan beberapa data set yang diambil dari UCI data repository. Hasil ujicoba menunjukkan bahwa algoritma modes linkage menghasilkan klaster yang stabil dan mempunyai rasio keakuratan yang lebih baik daripada k-modes.

Kata kunci: klasterisasi, data kategorikal, centroid linkage, k-modes

1. Pendahuluan

Salah satu operasi penting dalam data mining adalah melakukan klasterisasi terhadap data set berukuran besar. Klasterisasi merupakan teknik yang bisa digunakan untuk mengelompokkan obyek, dimana dalam satu kelompok yang terbentuk memiliki nilai keserupaan obyek yang besar dan antar kelompok memiliki nilai keserupaan yang kecil.

Akhir-akhir ini, penelitian dibidang klasterisasi data kategorikal sudah mulai berkembang, tetapi masih jauh lebih sedikit bila dibandingkan dengan klasterisasi pada tipe data numerik [5]. Banyak data yang direpresentasikan dalam bentuk data kategorikal, dimana nilai-nilai atributnya secara alami tidak bisa diperlakukan sebagai data numerik. Sebagai contoh atribut data kategorikal adalah atribut berdomain bentuk yang nilainya bisa berupa lingkaran, elips, bujur sangkar, dll. Dengan sifat atribut data kategorikal khusus seperti itu maka untuk mengklaster akan menjadi sulit dan kompleks dari pada mengklaster data numerik.

Dua pendekatan utama klasterisasi adalah pendekatan partisi dan pendekatan hirarki. Klasterisasi dengan pendekatan partisi atau sering disebut dengan *partition-based clustering* mengelompokkan data dengan memilah-milah data yang dianalisa ke dalam cluster-cluster yang ada. Klasterisasi dengan pendekatan hirarki atau sering disebut dengan *hierarchical clustering* mengelompokkan data dengan membuat suatu hirarki berupa dendogram dimana data yang mirip akan ditempatkan pada hirarki yang berdekatan dan yang tidak pada hirarki yang berjauhan. Di samping kedua pendekatan tersebut, ada juga clustering dengan pendekatan *automatic mapping* (Self-Organising Map/SOM) [1].

Terdapat berbagai macam algoritma klasterisasi. Beberapa algoritma dikhususkan hanya untuk tipe data tertentu, misalnya algoritma klasterisasi yang diperuntukkan mengolah data numerik seperti K-Means dan single linkage, complete linkage, centroid linkage, dan average linkage[6]. Terdapat juga algoritma yang dikhususkan untuk mengklaster data kategorikal seperti algoritma LIMBO [3], k-modes [10] dan berbagai algoritma pengembangan dari k-modes [2], [4],[5],[6],[8],[9].

Algoritma klasterisasi data kategorikal yang terkenal adalah algoritma yang dikembangkan oleh Huang [10], algoritma k-modes, merupakan pengembangan algoritma k-means agar dapat digunakan untuk mengklaster data kategorikal. Kontribusi utama dari algoritma k-modes ini adalah penggunaan ukuran ketidakserupaan berupa kecocokan suatu nilai atribut tiap dimensi terhadap titik pusat klaster, dimana titik pusat means diganti dengan modes, dan penggunaan metode yang didasarkan pada frekuensi untuk memutakhirkan modes.

Oleh karena algoritma k-modes yang dikembangkan oleh Huang menggunakan paradigma algoritma k-means maka algoritma k-modes yang dikembangkan mempunyai karakteristik yang sama dengan algoritma k-means. Dalam konteks ini hasil klaster dipengaruhi oleh nilai pembangkitan titik pusat awal klaster yang dipilih secara random sehingga klaster yang dihasilkan tidak selalu sama. Penelitian yang dilakukan oleh Funderlic [4], menemukan bahwa algoritma k-modes juga dapat terjebak pada optima lokal. Selain itu, hasil klasterisasi dengan algoritma k-modes juga menghasilkan nilai keserupaan dalam satu

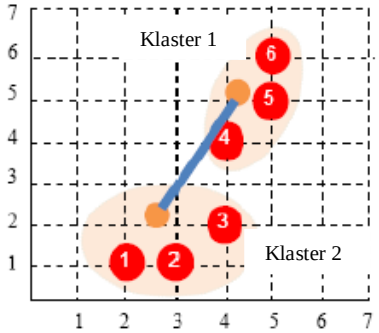
klaster yang rendah karena hanya mempertimbangkan penggunaan ukuran ketidakserupaan (jarak) berupa metode yang didasarkan pada frekuensi untuk memutakhirkan titik pusat klaster.

2. Algoritma Centroid Linkage

Algoritma centroid linkage merupakan algoritma klasterisasi hirarki. Penggabungan klaster pada centroid linkage didasarkan pada lokasi titik pusat (centroid) yang terbentuk pada tahap sebelumnya. Metode ini dibangun dengan memperhatikan pengecilan nilai standar deviasi klaster sekecil-kecilnya. Metode ini menggabungkan dua klaster melalui jarak terdekat diantara titik pusat antar klaster. Metode ini sangat ampuh untuk memperkecil variance within cluster karena melibatkan titik pusat pada saat penggabungan antar klaster. Metode ini juga baik untuk data yang mengandung outlier [6].

Gambar 1. Adalah ilustrasi dari hasil klaster algoritma centroid linkage. Jarak klaster 1 dan klaster 2 sama dengan jarak antara centroid kalster 1 dan centroid klaster 2. Algoritma centroid linkage selengkapnya adalah sebagai berikut:

1. Diasumsikan setiap data dianggap sebagai klaster. Jika n=jumlah data dan k=jumlah cluster, maka k=n.
2. Menghitung jarak antar klaster dengan Euclidian distance.
3. Pilih dua buah klaster yang mempunyai jarak centroid yang paling minimal dan gabungkan kedalam satu klaster baru (sehingga k=k-1). Lakukan pemutakhiran centroid tiap kali terjadi penggabungan klaster kemudian hitung jarak antar klaster.
4. Kembali ke langkah 3, dan ulangi sampai dicapai jumlah klaster yang diinginkan.



Gambar 1. Ilustrasi algoritma centroid linkage

3. Algoritma Klasterisasi Modes Linkage

Algoritma Modes Linkage adalah algoritma klasterisasi untuk data kategorikal yang mempunyai operasi utama sama dengan algoritma centroid linkage. Terdapat tiga modifikasi utama terhadap centroid linkage yaitu pemilihan rumus jarak, mengganti centroid dengan modes dan metode perhitungan frekuensi kemunculan data untuk memutakhirkan modes.

Sama dengan Huang [10], diasumsikan himpunan obyek yang akan diklaster tersimpan dalam dataset D yang didefinisikan dengan himpunan atribut $A_1,...,A_m$ dengan domain $D_1,...,D_m$. Domain D_i didefinisikan kategorikal jika bernilai terbatas dan tidak berurutan, sedemikian hingga hanya operasi perbandingan yang bisa dilakukan pada D_i. Sehingga untuk tiap nilai $a,b \in D_i$ maka $a=b$ atau $a \neq b$. Selajutnya akan dibahas mengenai ukuran ketidakserupaan, cara menentukan titik pusat dan cara memutakhirkan titik pusat.

3.1. Ukuran ketidakserupaan

Dinotasikan X dan Y adalah dua obyek kategorikal yang dideskripsikan oleh sejumlah m atribut. Pengukuran jarak antara X dan Y didefinisikan sebagai jumlah total ketidakcocokan atribute kategori-kategori yang berkorespondensi pada dua obyek tersebut. Semakin sedikit jumlah ketidakcocokan, semakin mirip dua obyek tersebut. Pengukuran keserupaan yang dimaksud adalah sebagai berikut:

$$d(X,Y)=\sum_{j=1}^{m}\delta(x_j,y_j) \quad (1)$$

dimana

$$\delta(x_j,y_j)=\begin{cases}0 & x_j=y_j\\ 1 & x_j\neq y_j\end{cases} \quad (2)$$

3.2. Himpunan mode

Anggap X adalah himpunan obyek kategorikal yang dideskripsikan oleh atribut kategorikal $A_1,A_2,...,A_m$. Maka dapat didefinisikan bahwa mode dari X adalah vektor $Q=[q_1,q_2...,q_m \in \Omega]$ yang meminimalkan

$$D(Q,X)=\sum_{i=1}^{n}d(X_i,Q) \quad (3)$$

dengan $X=\{X_1,X_2...,X_n\}$ dan d merupakan persamaan (2). Nilai Q tidak harus elemen X.

3.3. Menemukan himpunan mode

Selanjutnya cara menemukan himpunan mode adalah sebagai berikut. Anggap $n_{c_{k,j}}$ adalah jumlah

obyek yang mempunyai kategori $c_{k,j}$ pada atribut A_j dan $fr(A_j = c_{kj} \mid X) = \frac{n_{c_{kj}}}{n}$ adalah frekuensi relatif kategori $c_{k,j}$ dalam X. Fungsi $D(Q, X)$ akan minimal jika $fr(A_j = q_j \mid X) \geq fr(A_j = c_{k,j} \mid X)$ untuk $q_j \neq c_{k,j}$ untuk semua j = 1…m.

3.4. Algoritma Modes Linkage

Cara kerja algoritma Modes Linkage mengikuti paradigma algoritma centroid linkage. Terdiri dari langkah-langkah sebagai berikut:

1. Diasumsikan setiap data dianggap sebagai klaster. Jika n=jumlah data dan k=jumlah cluster, maka c=n.
2. Menghitung jarak antar klaster dengan persamaan (2).
3. Pilih dua buah klaster yang mempunyai jarak modes yang paling minimal dan gabungkan kedalam satu klaster baru (sehingga k=k-1). Lakukan pemutakhiran centroid tiap kali terjadi penggabungan klaster kemudian hitung jarak antar klaster dengan persamaan (2).
4. Kembali ke langkah 3, dan ulangi sampai dicapai jumlah klaster yang diinginkan.

4. Skenario Ujicoba

Di bagian ini akan dijelaskan beberapa dataset yang akan digunakan dan parameter ujicoba yang akan dilakukan. Algoritma k-modes [10] digunakan sebagai algoritma pembanding.

4.1. Dataset

Sebagai dataset masukan digunakan beberapa dataset dengan karakteristik sebagaimana ditunjukkan padaTabel 1. Semua dataset tersebut diambil dari UCI Learning Machine Repository [7]. Dataset Soybean Disease adalah data set standar yang sering digunakan dalam mengevaluasi metode-metode klasterisasi secara konseptual.

Tabel 1. Karakteristik dataset ujicoba

Nama dataset	Jumlah Kelas	Jumlah Atribut	Jumlah Data
Soybean	4	34	47
Hepatitis	2	13	155
Breast Cancer	2	9	286
Hayesroth	3	4	132
Balance scale	3	4	625

4.2. Parameter Eksperimen

Secara detail analisa yang dilakukan adalah berikut:

1. Keakuratan hasil klasterisasi
   Untuk menghitung keakuratan hasil klaster digunakan persamaan:

$$r = \frac{\sum_{i=1}^{k} a_i}{n} \quad (4)$$

   r = rasio akurasi hasil klaster
   n = jumlah data dalam data set
   ai = jumlah data pada klaster yang benar
2. Kecepatan waktu komputasi
   Membandingkan waktu yang diperlukan dalam menjalankan algortima modes linkage dan k-modes.

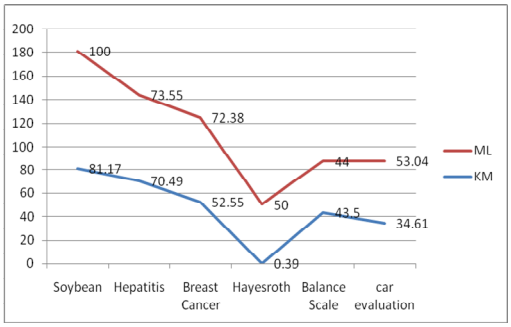
5. Hasil Ujicoba

Untuk mengukur keakuratan hasil dan waktu komputasi, maka untuk semua dataset tiap algoritma dijalankan sebanyak 20 kali. Hasil ujicoba dapat dilihat pada Tabel 2. KM adalah singkatan dari k-modes, sedang ML adalah singkatan dari modes linkage.

Tabel 2. Hasil ujicoba

Dataset	Rasio akurasi(%)		Waktu(milidetik)	
	KM	ML	KM	ML
Soybean	81.17	100.00	16	227
Hepatitis	70.49	73.55	27	3833
Breast Cancer	52.55	72.38	51	12177
Hayesroth	0.39	50.00	16	487
Balance Scale	43.5	44.00	74	47616
car evaluation	34.61	53.04	1299	1486813

Hasil ujicoba yang ditunjukkan pada Tabel 2. Dan Gambar 2. menunjukkan bahwa algoritma modes linkage menghasilkan tingkat akurasi yang lebih tinggi jika dibandingkan dengan algoritma k-modes pada semua dataset ujicoba. Namun demikian waktu komputasi yang dibutuhkan untuk menjalankan algoritma modes linkage lebih lama daripada k-modes. Semakin banyak jumlah data dan jumlah attribut maka waktu komputasi menjadi semakin besar.



Gambar 2. Grafik rasio akurasi antara k-modes dan modes linkage

6. Simpulan

Dari hasil ujicoba dan analisa dapat disimpulkan bahwa:

1. Jika dibandingan dengan algoritma K-Modes [3], algoritma modes linkage dapat menghasilkan hasil klaster yang lebih baik dengan nilai rasio akurasi yang dapat mencapai 100%.
2. Algoritma modes linkage membutuhkan waktu komputasi yang lebih lama dibandingkan dengan algoritma k-modes.

Daftar Pustaka

[1] Moore, 2001, A: K-means and Hierarchical Clustering – Tutorial Slides, November. Tutorial ada di alamat http://www-2.cs.cmu.edu/~awm/tutorials/kmeans.html.

[2] O. San, V. Huynh, Y. Nakamori, An alternative extension of the k-means algorithm for clustering categorical data. Int. J. Appl. Math. Comput. Sci., Vol. 14, No. 2, 241-247, 2004.

[3] P. Andritsos, P. Tsaparas, R. J. Miller, and K. C. Sevcik. Limbo: A linear algorithm to cluster categorical data. Technical report, UofT, Dept of CS, CSRG-467, 2003.

[4] R. E. Funderlic , M. T. Chu , N. Orlowski , D. Schlorff , J. Blevins , D., Convergence and Other Aspects of the K-Modes Algorithm for Clustering Categorical data, N.C. State University Department of Computer Science March 25. , 2004.

[5] S. Aranganayagi, and K. Thangavel, Improved K-Modes for Categorical Clustering Using Weighted Dissimilarity Measure, International Journal of Information and Mathematical Sciences 5:2, 2009.

[6] T. Karlita, A. Djunaidy, Algoritma Perbaikan Penentuan Titik Pusat Awal Berbasis Hirarki Untuk Klasterisasi Data Kategorikal, Seminar Pasca Sarjana, ITS Surabaya, 2006.

[7] UCI, Machine Learning Repository. (http://archive.ics.uci.edu/ml/datasets.html)

[8] Y. Sun, Q. Zhu, Z. Chen, An iterative initial-points refinement algorithm for categorical data clustering, Pattern Recognition Letters, 875-884, 23 July, 2002.

[9] Z. He, S. Deng, X. Xu, Improving K-Modes algorithm considering frequencies of attribute values in mode, International Conference on Intelligent Computing, China, 2005.

[10] Z. Huang, A fast clustering algorithm to cluster very large categorical data sets in data mining, In: SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, pp. 8, Tucson, AZ. 1997.