PENGKLASTERAN DATA DENGAN MENGGUNAKAN METODE MONOTETIS (STUDI KASUS PADA DATA KELUARGA)

Kania Sawitri

Jurusan Teknik Elektro ITENAS
email : kania@itenas.ac.id

Abstrak. Pengelompokan pada data variabel biner diperlakukan secara khusus. Ide dasarnya adalah untuk memilih salah satu variabel dan membagi himpunan dari objek ke dalam objek-objek dengan dan tanpa berhubungan dengan atribut. Dalam masing-masing dua himpunan bagian, salah satu variabel yang tersisa dipilih dan digunakan dengan cara yang sama untuk membagi bagian ini menjadi dua kelompok yang lebih kecil. Proses dilanjutkan sampai salah satu subhimpunan berisi objek tunggal (singleton) atau sampai variabel yang tersisa tidak dapat dipisahkan kedalam subhimpunan yang lain [2]. Oleh karena himpunan dari objek dibagi menjadi subhimpunan dan proses ini dilanjutkan terus di dalam setiap subhimpunan, maka algoritma ini termasuk algoritma divisif. Selain itu, karena setiap pemisahan dilakukan menggunakan variabel tunggal, maka metodenya dinamakan monotetik. Bagian paling penting dari algoritma ini adalah bagaimana memilih variabel untuk dipisahkan dalam sebuah subhimpunan. Idenya adalah untuk memilih variabel yang jumlah similaritas (kemiripan) untuk semua variabel lain adalah sebesar mungkin. Ukuran yang dapat digunakan dalam pengukuran variabel biner adalah ukuran kemiripan/ukuran asosiasi. Pada artikel ini dibahas hasil kajian teoritis tentang pengklasteran data dengan menggunakan metode monotetis. Hasil kajian teoritis tersebut diterapkan pada sebuah data yaitu data keluarga (studi kasus pada 10 data biner untuk 8 orang (lihat [2])).

Kata kunci : *Metode Monotetis, Data Biner, Ukuran Asosiasi.*

1. PENDAHULUAN

Analisis statistik multivariat merupakan metode statistika yang memungkinkan dilakukan penelitian terhadap lebih dari dua variabel secara bersamaan. Dengan menggunakan teknik analisis ini maka dapat dianalisis pengaruh beberapa variabel terhadap variabel-variabel lainnya dalam waktu yang bersamaan. Analisis multivariat digunakan karena pada kenyataannya masalah yang terjadi tidak dapat diselesaikan dengan hanya menghubung-hubungkan dua variabel atau melihat pengaruh satu variabel terhadap variabel lainnya [3].

Teknik analisis multivariat secara dasar diklasifikasi menjadi dua, yaitu analisis dependensi dan analisis interdependensi. Analisis dependensi berfungsi untuk menerangkan atau memprediksi variabel terikat dengan menggunakan dua atau lebih variabel bebas. Yang termasuk dalam klasifikasi ini adalah analisis regresi linier berganda, analisis diskriminan, analisis multivariat varian (MANOVA), analisis *conjoint* dan analisis korelasi kanonikal. Analisis interdependensi berfungsi untuk memberikan makna terhadap seperangkat variabel atau membuat kelompok-kelompok secara bersama-sama. Salah satu yang termasuk dalam klasifikasi ini adalah analisis *klaster* [3].

Analisis *klaster* merupakan suatu teknik multivariat yang digunakan untuk mengelompokkan objek pengamatan menjadi beberapa kelompok berdasarkan kemiripan antar objek, sehingga objek-objek yang berada dalam satu kelompok memiliki kemiripan yang lebih besar dibandingkan objek dari kelompok yang berbeda [1]. Dengan kata lain, analisis ini akan mengelompokkan objek sehingga setiap objek yang memiliki kesamaan dengan objek lainnya akan berada dalam kelompok yang sama. Setiap pengamatan harus memiliki homogenitas yang

tinggi dalam sebuah kelompok dan memiliki heterogenitas yang tinggi dengan kelompok yang lainnya [4].

Pengelompokan untuk data variabel biner diperlakukan secara khusus. Ide dasar pada pengelompokan data variabel biner adalah untuk memilih salah satu variabel dan membagi himpunan dari objek ke dalam objek-objek dengan dan tanpa berhubungan dengan atribut. Dalam masing-masing dua himpunan bagian, salah satu variabel yang tersisa dipilih dan digunakan dengan cara yang sama untuk membagi bagian ini menjadi dua kelompok yang lebih kecil (Perhatikan bahwa tidak perlu untuk menggunakan variabel kedua yang sama di kedua subhimpunan). Proses dilanjutkan sampai salah satu subhimpunan berisi objek tunggal (singleton) atau sampai variabel yang tersisa tidak dapat dipisahkan kedalam subhimpunan yang lain. Situasi terakhir ini hanya terjadi jika masing-masing variabel tetap konstan untuk semua objek dari subhimpunan ini [2].

Oleh karena himpunan dari objek dibagi menjadi subhimpunan dan proses ini dilanjutkan terus di dalam setiap subhimpunan, maka algoritma ini termasuk algoritma hirarkis. Lebih tepatnya termasuk algoritma divisif. Selain itu, karena setiap pemisahan dilakukan menggunakan variabel tunggal, maka metodenya dinamakan monotetis.

Bagian paling penting dari algoritma ini adalah bagaimana memilih variabel untuk dipisahkan dalam sebuah subhimpunan. Idenya adalah untuk memilih variabel yang jumlah similaritas (kemiripan) untuk semua variabel lain adalah sebesar mungkin. Ukuran yang dapat digunakan adalah ukuran kemiripan. Ukuran kemiripan yang sering digunakan dalam pengukuran variabel biner adalah ukuran asosiasi.

Peneliti biasanya memakai beberapa peubah untuk menggambarkan karakteristik populasi yang diteliti. Seringkali banyak penelitian melibatkan data dengan skala pengukuran seperti numerik (interval, rasio) dan kategorik (nominal, ordinal). Dalam penelitian ini akan digunakan data dengan skala pengukuran kategorik nominal dengan dua keadaan, jenis data ini dinamakan data biner. Variabel biner hanya memiliki dua kemungkinan hasil atau keadaan. Misalnya, ketika mengelompokkan orang beberapa variabel biner dapat digunakan: pria/wanita, perokok/bukan perokok, menjawab ya/tidak untuk pertanyaan tertentu, dan sebagainya. Dalam matriks data, variabel seperti itu sering dikodekan sebagai nol atau satu. Seringkali 1 dimaksudkan untuk menjelaskan bahwa atribut tertentu hadir, misalnya merokok, sedangkan 0 menunjukkan tidak merokok.

Pada artikel ini dibahas hasil kajian teoritis tentang pengklasteran data dengan menggunakan metode monotetis. Hasil kajian teoritis tersebut diterapkan pada sebuah data yaitu data keluarga (studi kasus pada 10 data biner untuk 8 orang (lihat [2])).

2. METODE MONOTETIS

Algoritma monotetis atau *Monothetic analysis algorithms* disingkat algoritma MONA [2], dimaksudkan untuk data yang terdiri dari variabel biner. Variabel biner hanya memiliki dua kemungkinan hasil atau keadaan: 1 jika atribut tertentu hadir dan 0 jika tidak ada. Misalnya, ketika mengelompokkan orang beberapa variabel biner dapat digunakan: pria/wanita, perokok/bukan perokok, menjawab ya/tidak untuk pertanyaan tertentu, dan sebagainya. Dalam matriks data, variabel seperti itu sering dikodekan sebagai nol atau satu. Ketika variabel f adalah biner, objek i akan memiliki salah satu dari $x_{if} = 0$ atau $x_{if} = 1$. Seringkali 1 dimaksudkan untuk menjelaskan bahwa atribut tertentu hadir (misalnya, merokok), sedangkan 0 menunjukkan tidak merokok.

Algoritma MONA termasuk algoritma divisif dan memanfaatkan sepenuhnya aspek biner. Pada setiap langkah salah satu variabel yang digunakan untuk membagi data, dengan memisahkan objek yang variabel mengambil nilai 1 dari objek yang mengambil nilai 0. Untuk menentukan

variabel yang digunakan, pertamakali dihitung ukuran hubungan antara variabel (untuk alasan ini, algoritma ini sering disebut "analisis asosiasi") dan kemudian memilih variabel yang paling representatif atau yang paling kuat terkait dengan variabel lain.

Ukuran Asosiasi

Setelah matriks data lengkap diperoleh, algoritma MONA mulai dengan pengelompokan yang sebenarnya. Algoritma yang digunakan adalah hirarkis jenis divisif. Pada setiap langkah pemisahan, dipilih satu variabel yang digunakan untuk membagi himpunan, dengan memisahkan objek yang variabel mengambil nilai 1 dari objek yang variabel mengambil nilai nol. Pemilihan variabel didasarkan pada hubungan antara variabel, dengan ukuran asosiasi antara variabel f dan g dihitung sebagai berikut:

$$A_{fg} = |a_{fg}d_{fg} - b_{fg}c_{fg}|$$

di mana nilai-nilai a_{fg}, b_{fg}, c_{fg}, dan d_{fg} diperoleh dari tabel kontingensi berikut:

		Variabel g	
		1	0
Variabel f	1	a_{fg}	b_{fg}
	0	c_{fg}	d_{fg}

Perhatikan bahwa $a_{fg} + b_{fg} + c_{fg} + d_{fg}$ adalah jumlah total dari variabel f.

Untuk setiap variabel f, ukuran asosiasi dengan semua variabel lain ditambahkan dan diperoleh ukuran keseluruhan sebagai berikut:

$$A_f = \sum_{g \neq f} A_{fg}$$

Jumlah maksimum dari f adalah :

$$A_t = \max_f A_f$$

Sekali lagi, jika nilai maksimal yang sama ditemukan untuk beberapa variabel, t dipilih sebagai salah satu yang muncul pertama kali.

Asosiasi antara variabel dapat dianggap sebagai semacam kesamaan di antara mereka. Asosiasi antara variabel dan dirinya sendiri biasanya sangat besar, karena nilai b_{ff} dan c_{ff} keduanya 0. Perhatikan bahwa asosiasi antara variabel dan dirinya sendiri dapat bernilai 0 dalam kasus yang ekstrim baik a_{ff} atau d_{ff} nilainya 0. Namun, dalam hal ini f tidak dapat digunakan untuk memisahkan himpunan dari objek. Di sisi lain, jika dua variabel memberikan informasi yang sangat berbeda, hubungan mereka sangat kecil. Sebenarnya, ukuran asosiasi ini bukan kesamaan yang sebenarnya karena dapat mengambil nilai-nilai yang melebihi 1.

Variabel yang dipilih untuk pemisahan adalah salah satu variabel yang asosiasi keseluruhan dengan semua variabel lainnya adalah yang terbesar. Oleh karena itu dapat dikatakan bahwa total ketidaksamaan dengan variabel lain adalah sekecil mungkin. Dengan demikian variabel ini dapat dianggap sebagai medoid variabel.

Setelah pemisahan pertama dari objek menjadi dua himpunan bagian, proses ini diulang dalam masing-masing himpunan bagian. Variabel yang digunakan untuk pemisahan pertama tidak perlu digunakan kembali, jadi pemisahan berikutnya menggunakan variabel yang tersisa, karena dibutuhkan nilai yang identik untuk semua objek dari himpunan bagian yang sama. Pemisahan masing-masing bagian dilakukan secara independen.

Langkah-langkah pemisahan dilanjutkan sampai himpunan bagian terdiri dari satu objek atau sampai variabel yang tersisa tidak dapat digunakan untuk pemisahan (karena semua objek

memiliki nilai yang sama untuk variabel-variabel ini). Dalam situasi yang terakhir himpunan bagian disimpan secara keseluruhan.

Langkah pemisahan sangat bergantung pada asosiasi antara variabel yang tersisa. Cara di mana asosiasi ini dihitung mengakibatkan bahwa hanya sejumlah nilai yang berbeda dapat diperoleh. Hal ini terutama terjadi selama tahap terakhir dari algoritma. Sebagian besar aplikasi situasi yang sama terjadi ketika hanya ada variabel dan atau objek yang tersisa. Secara khusus, hal ini terjadi ketika hanya ada dua variabel yang tersisa untuk pemisahan himpunan bagian.

3. HASIL PENERAPAN METODE MONOTETIS

Untuk memberikan ilustrasi hasil penerapan kajian teoritis metode monotetis, berikut ini diberikan contoh penerapannya pada data keluarga (studi kasus pada 10 data biner untuk 8 orang (lihat [2])). Oleh karena itu akan diambil 10 variabel biner yang diukur pada 8 orang (objek) untuk diolah sebagai contoh dalam menerapkan metode monotetis, dengan sepuluh variabel biner yang terlibat adalah sebagai berikut : V_B=Jenis kelamin (Pria=1, wanita=0); V_C= Menikah (ya=1, tidak=0); V_D= Rambut (Pirang=1, hitam=0); V_E = Mata (Biru=1, coklat=0); V_F = Berkacamata (ya=1, tidak=0); V_G = Wajah (Bulat=1, oval=0); V_H = (Pesimis=1, optimis=0); V_I = (malam=1, siang=0); V_J = Punya anak (ya=1, tidak=0); V_K = Tangan (kanan=1, kiri=0).

Tabel 1. Data Keluarga dengan 10 variabel Biner untuk 8 Orang

Orang	Jenis Kelamin (Pria=1 Wanita=0)	Menikah Ya=1 No=0	Rambut Pirang=1 Hitam=0	Mata Biru=1 Coklat=0	Berkaca Mata Ya=1 No=0	Wajah Bulat=1 Oval=0	Pesimis=1 Optimis=0	Malam=1 Siang=0	Punya Anak Ya=1 No=0	Tangan Kanan=1 Kiri=0
Ilan	1	0	1	1	0	0	1	0	0	0
Jacqueline	0	1	0	0	1	0	0	0	0	0
Kim	0	0	1	0	0	0	1	0	0	1
Lieve	0	1	0	0	0	0	0	1	1	0
Leon	1	1	0	0	1	1	0	1	1	0
Peter	1	1	0	0	1	0	1	1	0	0
Talia	0	0	0	1	0	1	0	0	0	0
Tina	0	0	0	1	0	1	0	0	0	0

Data ini terdiri dari 10 variabel biner yang diukur pada 8 orang (objek). Hanya 0 dan 1 yang muncul dalam data, sehingga tidak ada pengukuran yang hilang (nilai yang berbeda dari 0 dan 1).

Analisis dilakukan pada data ini menggunakan semua variabel. Pada langkah pertama dari perhitungan ‘variabel menikah’ terpilih sebagai perwakilan terbesar, dan oleh karena itu yang paling memadai untuk membagi kumpulan data menjadi dua kelompok. Objek pertama dari klaster adalah salah satu yang muncul pertama kali dalam matriks data, maka semua objek yang diberikan yang memiliki nilai yang sama untuk variabel pemisahan. Orang pertama (objek) adalah ILA, yang belum menikah. Dia diikuti oleh KIM, TAL, dan TIN. Setelah klaster pertama

ini, objek yang diberikan yang memiliki nilai yang berlawanan adalah JAC, LIE, LEO, dan PET. Dalam kedua kelompok objek muncul dalam urutan yang sama seperti dalam matriks data.

Sangat menarik untuk diamati bahwa dalam penerapan ini variabel yang digunakan untuk divisi pertama adalah secara intuitif yang paling penting, karena membedakan antara anak-anak dan orang dewasa.

Pada langkah kedua, kedua kelompok dipisahkan menggunakan salah satu variabel yang tersisa. Klaster anak dipisahkan berdasarkan 'variabel warna rambut', sedangkan untuk orang dewasa berdasarkan 'variabel jenis kelamin' adalah yang paling terletak terpusat.

Pada langkah ketiga klaster berikut diperiksa: {ILA, KIM}, {TAL, TIN}, {JAC, LIE}, dan {LEO, PET}. Yang pertama dari kelompok ini hanya dapat dipisahkan dengan menggunakan 'variabel jenis kelamin' dan 'tangan kanan'. Variabel lain mengambil nilai yang sama untuk anak-anak, sehingga mereka tidak akan lagi dianggap. Kesamaan antara variabel 'jenis kelamin' dan 'tangan kanan' akan dihitung dengan menggunakan tabel berikut:

		tangan kanan	
		1	0
Jenis kelamin	1	0	1
	0	1	0

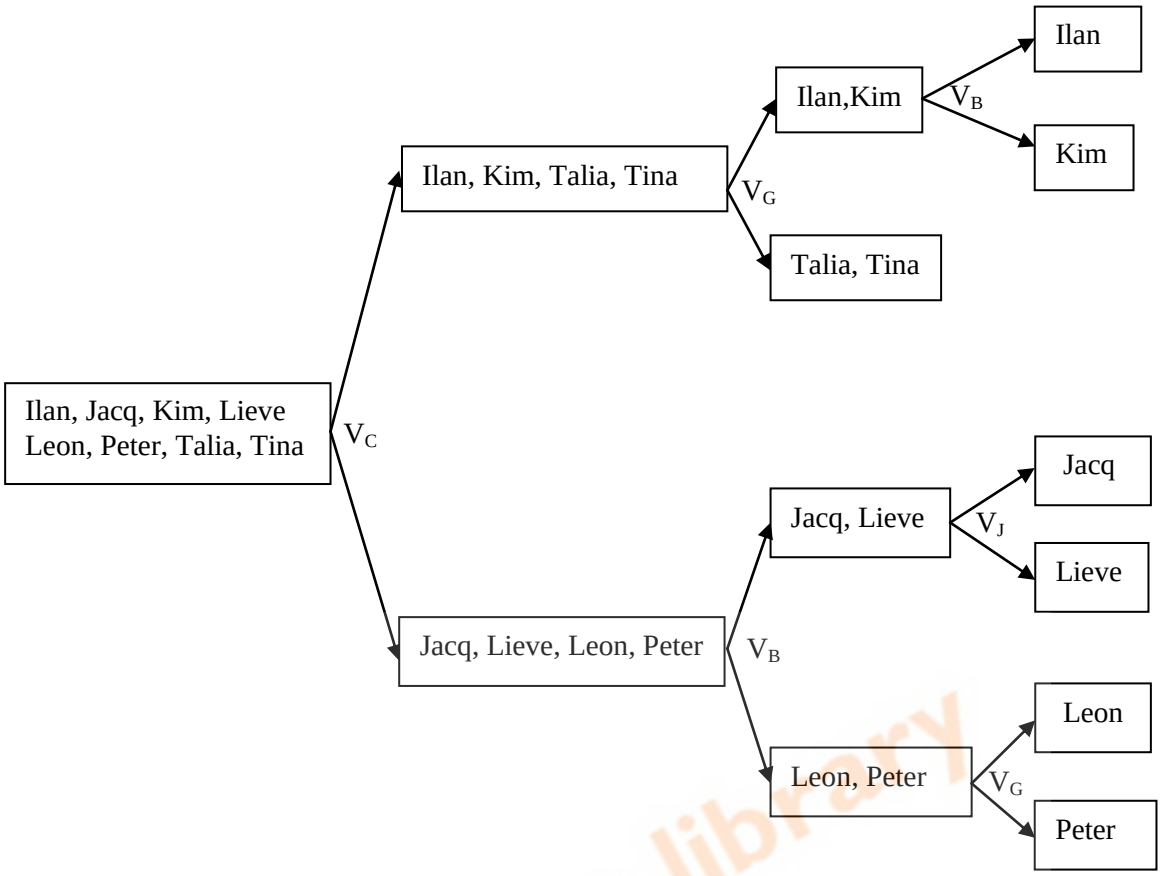
Kesamaan antara dua variabel sama dengan $|0 \times 0 - 1 \times 1| = 1$, yang juga merupakan kesamaan total untuk setiap variabel. Dalam situasi di mana beberapa variabel menghasilkan kesamaan total yang sama, salah satu yang muncul pertama kali dalam matriks data yang dipilih untuk pemisahan. Hal ini tentu saja sembarang pilihan. Perhatikan bahwa ketika ada dua objek dalam sebuah klaster, variabel pertama dipilih yang mengambil nilai yang berbeda di dalamnya.

Klaster berikutnya yang harus dipertimbangkan dalam langkah 3 adalah {TAL, TIN}. Karena semua variabel yang tersisa mengambil nilai-nilai yang sama untuk dua anak ini, mereka tidak dapat dipisahkan. Dalam situasi seperti klaster ini tidak akan lagi dianggap dalam langkah-langkah berikutnya. Perhatikan bahwa objek klaster tersebut adalah identik untuk semua variabel dari kumpulan data, tidak hanya untuk orang-orang yang tersisa. Mereka berada diklaster yang sama karena mereka juga sama untuk variabel yang digunakan dalam mendapatkan klaster ini.

Klaster tersisa {JAC, LIE} dan {LEO, PET} dipisahkan oleh variabel 'berkacamata' dan 'wajah', karena ini adalah variabel pertama dimana obyek memiliki nilai yang berbeda. Dalam hal ini struktur hirarkis yang dihasilkan oleh algoritma dapat disimpulkan dari hasil akhir. Ini dimulai dengan urutan akhir dari objek, yang dapat berbeda dari yang asli.

Hasil akhirnya dirangkum sebagai berikut: pemisahan pertama, dengan menggunakan variabel 'menikah', membagi data ke dalam klaster {ILA, KIM, TAL, TIN} dan {JAC, LIE, LEO, PET}. Pada bagian pertama dari kelompok ini variabel 'rambut' digunakan untuk divisi berikutnya, dan bagian kedua adalah variabel 'jenis kelamin'. Variabel yang digunakan dalam langkah ketiga adalah kurang penting. Namun, perlu diketahui bahwa klaster (TAL, TIN} tidak dapat dipisahkan.

Hasil penerapan analisis monotetis pada data keluarga pada tabel 1 dapat dilihat pada Gambar 1 sebagai berikut:



Gambar 1. Hasil Pengklasteran Metode Monotetis pada Data Keluarga

4. KESIMPULAN

Hasil pengelompokan dengan menggunakan metode monotetis pada contoh data sampel: sepuluh variabel biner untuk delapan orang data keluarga memberikan hasil tujuh klaster. Pengelompokan kedelapan orang data keluarga ditunjukkan pada Gambar 1.

Keuntungan menggunakan algoritma MONA pada pengklasteran data biner adalah dalam setiap langkah berikutnya, setiap klaster dibagi dengan cara yang sama oleh salah satu variabel yang tersisa. Hal ini menggambarkan fitur bagus dari metode MONA, kita tidak hanya mendapatkan klaster, tetapi kita juga tahu variabel mana yang bertanggung jawab untuk pemisahan.

Algoritma MONA cukup primitif, tetapi pada sisi lain kesederhanaan metode ini memberikan beberapa keuntungan yakni hasilnya mudah ditafsirkan karena kita tahu variabel mana yang menyebabkan pengelompokan dan algoritmanya cepat serta mampu menangani himpunan data yang besar [2].

DAFTAR PUSTAKA

[1] Johnson, R.A. and Wincern, D.W. (1982). *Applied Multivariate Statistical Analysis.* Prentice Hal Inc, New Jersey.

[2] Kaufman, et. all. (2005). *Finding Groups in Data.* John Wiley & Sons, New Jersey.

[3] Narimawati, U.(2007). *Teknik-teknik Analisis Multivariat untuk Riset Ekonomi.* Graha Ilmu, Jakarta.

[4] Sharma, S.(1996). *Applied Multivariate Technique.* John Wiley & Sons, Canada.