ANALISIS KARAKTERISTIK KELOMPOK DENGAN MENGGUNAKAN PENDEKATAN *CLUSTER ENSEMBLE*

Dyah Paminta Rahayu(dyahp@ut.ac.id)
Jurusan Matematika FMIPA Universitas Terbuka

ABSTRAK

Pengelompokkan merupakan salah satu tehnik data mining yang digunakan untuk mengelompokkan data berdasarkan kemiripan atribut dari data obyek. Pada umumnya algoritma pengelompokan dikembangkan hanya untuk memproses salah satu tipe data kategori atau numerik. Tidak banyak algoritma yang dikembangkan untuk memproses data campuran kategori dan numerik. Salah satu algoritma untuk memproses data campuran adalah algCEBMDC, algoritma pengelompokan dengan pendekatan cluster ensemble. Tujuan penelitian ini adalah untuk menganalisis karakteristik hasil pengelompokan algoritma algCEBMDC. Metode penelitian mengikuti alur kerja data mining dan algoritma algCEBMDC. Data yang digunakan adalah data mahasiswa non aktif Program Studi Matematika FMIPA, Universitas Terbuka. Data awal bertipe campuran dibersihkan untuk mendapatkan data bersih siap proses, kemudian dipisah menjadi dua berdasarkan tipe datanya: kategori dan numerik. Data kategori diproses menggunakan algoritma QROCK, menghasilkan 44 kelompok yang diperoleh pada threshold 0.98 dengan nilai kohesi 2044. Data numerik diproses menggunakan algoritma AGNES, menghasilkan 69 kelompok yang diperoleh dari kombinasi ukuran jarak Cityblock *distance dan metode penggabungan* Average link *dengan nilai cophenet 0,822. Hasil dari kedua pengelompokan digabung, dianggap sebagai data kategori, kemudian diproses menggunakan algoritma QROCK. Kelompok-kelompok yang dihasilkan memiliki kesamaan karakteristik pada pendidikan akhir, status pekerjaan, status perkawinan, dan jenis kelamin. Faktor prestasi akademik menunjukkan bahwa tingkat kelulusan matakuliah dalam dua semester pertama sangat rendah. Dapat dikatakan bahwa dua semester pertama merupakan masa kritis bagi mahasiswa Program Studi Matematika UT.*

Kata kunci: algoritma algCEBMDC, kompleksitas, pengelompokan

ABSTRACT

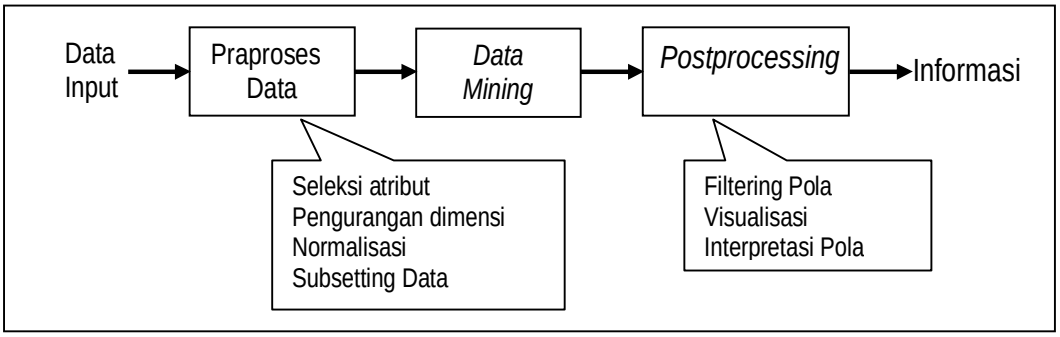
Clustering is one of data mining techniques which is used to group databased on similarity of the object data attributes. In general clustering algorithm is developed to process only one type of data, either category or numerical data type. Not many algorithms were developed to process the mixture between category and numerical data. One algorithm to process the mixed data is algCEBMDC, a clustering algorithm using cluster ensemble approach. The purpose of this study was to analyze the characteristics of the results of clustering algorithms algCEBMDC. The research methods follow the work-flow of data mining and algCEBMDC algorithm. The data used is the data of inactive students of Mathematics study program in Universitas Terbuka (The Indonesia Open University). First, the data is cleared up to get clean data ready for processing, and then is separated into two groups based on the type of category data and numerical data. The category data is processed using QROCK algorithm, producing 44 groups which is obtained at the 0.98 threshold value with

cohesion of 2044. The numerical data is processed using AGNES algorithm, generating 69 groups which is derived from a combination of Cityblock Distance and Average link method with cophenet value of 0,822. The results of the two grouping are combined, considered as a data category, then is processed using QROCK algorithm. The resulting groups had similar characteristics on the end of education, employment status, marital status, and gender. The academic achievement factors indicate that the passing level of courses in the first two semesters are very low. It can be concluded that the first two semesters is a critical time for distance education students in mathematic study program.

Keywords: algCEBMDC algorithm, characteristics, clustering

Sebagai Universitas yang tergolong dalam ”Mega University”, Universitas Terbuka (UT) memiliki data kemahasiswaan dengan jumlah yang sangat besar. Gudang data tersebut sebenarnya dapat dimanfaatkan oleh pengelola untuk mengembangkan institusi, misalnya untuk peningkatan efektifitas pemasaran atau pengurangan biaya operasional. Selain itu gudang data dapat juga digunakan untuk memecahkan masalah-masalah berikut: bagaimana mengelompokkan mahasiswa yang memiliki kesamaan karakteristik tertentu, mengestimisasi data yang hilang, meningkatkan performa akademik mahasiswa, atau mengurangi resiko kegagalan mahasiswa. Untuk dapat memanfaatkan gudang data, dibutuhkan suatu teknologi yang dapat dengan cepat menganalisis data dalam jumlah besar. Teknologi yang dimaksud adalah *data mining*.

Data mining adalah eksplorasi dan analisis secara otomatis atau semi otomatis terhadap data besar dengan tujuan untuk menemukan pola baru dan bermakna yang mungkin masih belum diketahui (Tan *et al.* 2006). *Data mining* merupakan bagian integral dari *Knowledge Discovery in Databases* (KDD). Keseluruhan proses KDD, mulai dari data masukan sampai menjadi informasi ditunjukkan oleh Gambar 1.



Gambar 1. Proses k*nowledge discovery in databases* (Tan *et al*, 2006)

Chong (2010) memanfaatkan beberapa teknik *data mining*, yaitu *classification trees*, *multivariate adaptive regression splines* (MARS), and *neural networks* untuk menganalisis *student retention* pada Arizona State University (ASU). Saxena (2002) menggunakan salah satu teknik *data mining*, yaitu analisis pengelompokkan dengan pendekatan *hierarchical clustering* untuk menganalisis data mahasiswa *India Open University,* sedangkan Sheela (2010) menggunakan metode pengelompokkan K-means untuk menemukan *knowledge* dari data akademik mahasiswa *Department of Computer Science, University of Agriculture, Faisalabad*.

Pada analisis pengelompokkan (*cluster analysis*) data, dilakukan pengelompokkan berdasarkan kemiripan atribut dari data obyek. Dalam hal ini data obyek yang berada di dalam kelompok yang sama memiliki kemiripan satu sama lain. Sedangkan dengan data obyek di dalam kelompok lain sama sekali tidak memiliki kemiripan. Semakin besar tingkat kemiripan antar obyek di dalam kelompok dan semakin besar tingkat perbedaan antar kelompok, berarti semakin baik pengelompokkan tersebut (Han & Kamber, 2001).

Pada umumnya algoritma pengelompokkan dikembangkan hanya untuk memproses salah satu tipe data kategori atau numerik. Tidak banyak algoritma yang dikembangkan untuk memproses data campuran kategori dan numerik. Padahal secara umum data riil memiliki atribut dengan tipe campuran (kategori dan numerik).

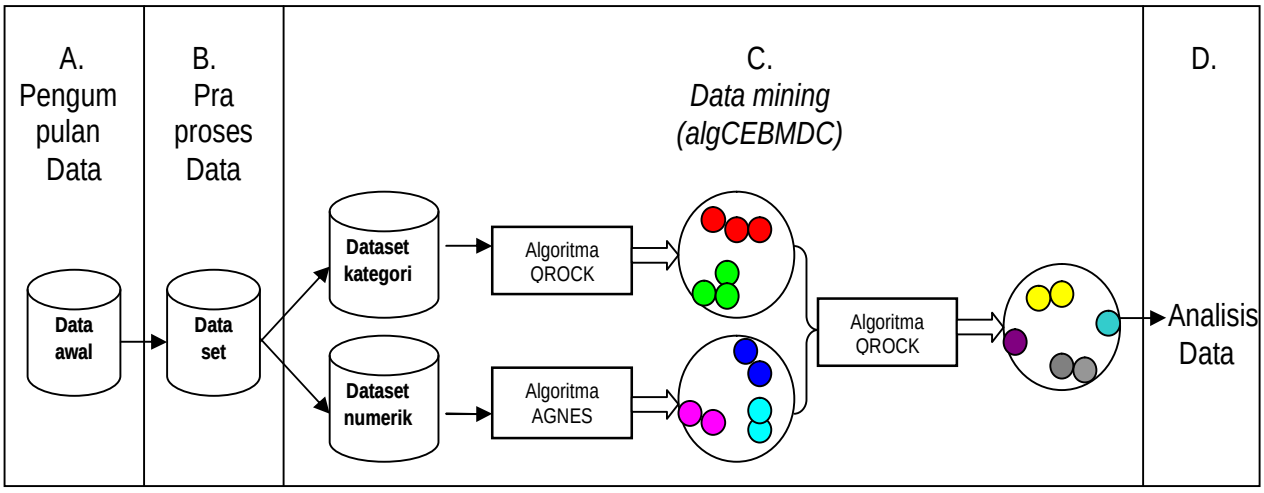
*AlgCEBMDC d*an *k-prototype* adalah dua contoh algoritma pengelompokkan yang bekerja pada data bertipe campuran. Zengyou (2002) membandingkan kedua algoritma tersebut dan menunjukkan bahwa *algCEBMDC* memiliki akurasi lebih baik dibandingkan *k-prototype*.

Algoritma *algCEBMDC* merupakan analisis pengelompokkan dengan pendekatan *cluster ensemble*. Algoritma ini menawarkan suatu teknik baru, yaitu teknik *devide-and-conquer*. Pertama, data awal bertipe campuran dipisah menjadi dua data kategori dan data numerik. Selanjutnya, kedua data tersebut diproses secara terpisah dengan menggunakan algoritma pengelompokkan yang sesuai dengan tipe masing-masing data. Untuk mendapatkan hasil akhir, kelompok-kelompok yang dihasilkan oleh kedua algoritma digabung dan dipandang sebagai data baru dengan tipe kategori, kemudian diproses dengan menggunakan algoritma pengelompokkan untuk data kategori (Zengyou, Xiaofe, & Shengchum. 2002).

Artikel ini menganalisis karakteristik hasil pengelompokkan algoritma *algCEBMDC* pada data mahasiswa nonaktif Program Studi Matematika FMIPA, Universitas Terbuka.

METODE

Metode penelitian dikembangkan berdasarkan alur proses KDD yang ditulis oleh Tan, Steinbach, dan Kumar (2006) dan algoritma *algCEBMDC* (Zengyou, Xiaofe, & Sheng, 2002). Skema metode penelitian disajikan oleh Gambar 2.



Gambar 2. Skema metode penelitian

Proses dimulai dengan pengumpulan data. Data awal yang berkaitan dengan demografi, latar belakang pendidikan, dan prestasi akademik mahasiswa diperoleh dari Pusat Komputer Universitas Terbuka pada tahun 2008. Data yang diperoleh berjumlah 5883 mahasiswa nonaktif Program Studi Matematika FMIPA-UT, memiliki 33 atribut campuran yang terdiri dari 23 atribut kategori dan 10 atribut numerik.

Praproses data yang dilakukan adalah pembersihan data, reduksi data, pemisahan data, dan transformasi data. Pembersihan data dilakukan karena data seringkali memiliki *record* dengan nilai atribut yang tidak lengkap, kosong, tidak konsisten, dan *noisy.* Data demikian harus dihapus, tidak disertakan dalam proses *data mining* karena dapat mempengaruhi hasil akhir secara negatif. Reduksi data yang dilakukan adalah seleksi atribut, karena tidak semua atribut relevan dengan kebutuhan penelitian. Transformasi diperlukan untuk mengkonversi data ke dalam format sesuai dengan kebutuhan.

Untuk mendapatkan hasil optimal, pemilihan algoritma pengelompokkan harus tepat karena kualitas hasil akhir dari suatu proses data mining tidak hanya tergantung pada kualitas data tetapi juga algoritma yang digunakan. Setiap algoritma pengelompokkan akan menghasilkan kelompok dengan tipenya masing-masing (Tan *et al*, 2006). Pada penelitian ini algoritma pengelompokkan yang digunakan adalah algoritma *algCEBMDC*.

Pengelompokkan data kategori menggunakan fungsi *just_qrock_edit*, ditulis oleh Marisa (2008) dengan menerapkan algoritma QROCK yang merupakan percepatan dari algoritma ROCK. Masukan dari fungsi ini adalah data kategori dan *treshold* sebagai ukuran kemiripan antar obyek. Untuk mendapatkan kelompok terbaik, kelompok yang dihasilkan dievaluasi menggunakan ukuran nilai kohesi. Semakin tinggi total nilai kohesi suatu hasil pengelompokkan, semakin baik kelompok yang dihasilkan.

Pengelompokkan data numerik menerapkan algoritma *AGNES* dengan menggunakan fungsi-fungsi yang tersedia dalam matlab 7.0, yaitu *pdist* untuk menghitung jarak antar obyek, *lingkage* untuk menggabungkan obyek atau kelompok, *chopenet* untuk menghitung nilai *chophenet,* dan *dendrogram* untuk membuat *dendrogram* dari kelompok yang terbentuk. Hasil pengelompokkan dievaluasi dengan cara menghitung nilai *cophenet*. Hasil pengelompokkan dikatakan baik jika nilai *cophenet* mendekati angka 1. Hasil pengelompokkan algoritma *QROCK* dan algoritma *AGNES* digabung dan dipandang sebagai data baru dengan tipe kategori kemudian diproses dengan menggunakan algoritma *QROCK.*

HASIL DAN PEMBAHASAN

Praproses Data

Pada proses pembersihan data ditemukan data yang memiliki atribut dengan nilai tidak lengkap, kosong, dan tidak konsisten. Nilai tidak lengkap terdapat pada atribut Tanggal Lahir pada bagian tahun lahir. Ketidaklengkapan pengisian tahun lahir mengakibatkan kesalahan dalam perhitungan umur mahasiswa karena umur dihitung berdasarkan tahun lahir. Nilai kosong terdapat pada atribut IPK (Indeks Prestasi Akademik) dan SKS (Satuan Kredit Semester). Hal tersebut bisa jadi karena belum ada satupun mata kuliah yang lulus, atau mahasiswa hanya melakukan pendaftaran sebagai mahasiswa UT tetapi tidak pernah mengikuti ujian. Nilai tidak konsisten terdapat pada atribut IPK dalam hubungannya dengan atribut SKS, yaitu terdapat 3 mahasiswa yang telah menempuh dan lulus beberapa mata kuliah tetapi IPK yang didapat 0. Nilai tidak konsisten juga terdapat pada atribut SKS dalam kaitannya dengan atribut Lama Studi, yaitu terdapat 5 mahasiswa yang rata-rata perolehan SKS tiap semester melebihi maksimum jumlah mata kuliah yang dapat

diambil tiap semesternya. Nilai tidak konsisten juga terdapat pada atribut Lama Studi, akibat dari kesalahan pada pengisian atribut Registrasi Akhir karena atribut Lama Studi dihitung berdasarkan atribut Registrasi Akhir. Terdapat 1751 *record* yang harus dihapus karena memiliki atribut dengan nilai tidak lengkap, kosong, dan tidak konsisten.

Reduksi data yang dilakukan dalam penelitian ini adalah seleksi atribut. Dari 33 atribut yang dimiliki, terdapat 23 atribut yang dihapus karena tidak relevan dengan kebutuhan penelitian. Atribut nama dan alamat mahasiswa merupakan contoh atribut yang dihapus karena walaupun atribut tersebut penting bagi mahasiswa tetapi tidak relevan dengan kebutuhan penelitian.

Setelah pembersihan dan reduksi data, *dataset* yang dianggap bersih dan siap diproses berjumlah 4132 *record*s dengan 10 atribut bertipe campuran: 6 atribut kategori dan 4 atribut numerik. Data tersebut diberi nama DataMhs dengan struktur sebagai berikut:

- Atribut Kategori. Semua status dalam atribut kategori dikonversi kedalam kode numerik. Atribut kategori meliputi:
    1. Jurusan Asal adalah atribut yang menerangkan jurusan dari pendidikan akhir yang dimiliki mahasiswa. Sebagai contoh, jika mahasiswa memiliki pendidikan akhir SLTA maka bisa jadi berasal dari jurusan IPA, IPS atau STM. Terdapat 79 (tujuh puluh sembilan) kode numerik untuk menerangkan status dari Jurusan Asal. Sebagai contoh '101' untuk 'SMTA Umum IPA/IPS'.
    2. UPBJJ adalah atribut yang menerangkan wilayah keberadaan mahasiswa. Terdapat 37 (tiga puluh tujuh) kode numerik untuk menerangkan status dari UPBJJ. Sebagai contoh '21' untuk 'UPBJJ UT Jakarta'.
    3. Pendidikan Akhir adalah atribut yang menerangkan pendidikan terakhir sebelum menjadi mahasiswa UT. Terdapat 6 (enam) kode numerik untuk menerangkan status dari Pendidikan Akhir, yaitu 1(SLTA), 2(D1), 3(D2), 4(D3), 5(S1), dan 6(S2).
    4. Status Kerja adalah atribut yang menerangkan jenis pekerjaan dari mahasiswa. Terdapat 5 (lima) kode numerik untuk menerangkan status dari Status Kerja, yaitu 2 (PNS), 3 (Swasta), 4 (Wiraswasta), 5 (Tidak Bekerja), dan 6 (Bekerja).
    5. Status Kawin memiliki 2(dua) kode numerik; 1(kawin) dan 0 (tidak kawin).
    6. Jenis Kelamin memiliki 2 (dua) kode numerik; 1 (Laki-laki) dan 0 (perempuan).
- Atribut Numerik meliputi:
    1. Umur adalah atribut yang menerangkan usia mahasiswa ketika pertama kali mendaftar sebagai mahasiswa UT, memiliki rentang nilai antara 16 sampai dengan 66 tahun.
    2. IPK adalah atribut yang menerangkan IPK yang diperoleh selama menjadi mahasiswa UT, memiliki rentang nilai antara 1 sampai dengan 4.
    3. SKS adalah atribut yang menerangkan jumlah SKS dari sejumlah mata kuliah yang sudah berhasil ditempuh dan lulus dengan nilai minimal D, memiliki rentang nilai antara 3 sampai dengan 175 SKS.
    4. Lama Studi adalah atribut yang menerangkan berapa semester mahasiswa mengikuti perkuliahan di UT, memiliki rentang nilai antara 1 sampai dengan 34 semester.

Sebelum proses *data mining,* DataMhs dipisah menjadi dua berdasarkan tipe dari atributnya. Data dengan atribut kategori diberi nama DataKategori dan data dengan atribut numerik diberi nama DataNumerik.

Beberapa atribut dari DataKategori memiliki nilai dengan kode numerik yang sama. Sebagai contoh: atribut Pendidikan Akhir 'D1' memiliki kode numerik yang sama dengan atribut Status Kerja

'PNS', yaitu '2', sedangkan atribut Status Kawin 'Tidak Kawin' memiliki kode numerik yang sama dengan atribut Jenis Kelamin 'Perempuan', yaitu '0'. Hal demikian dapat mengacaukan hasil perhitungan *similarity* antar obyek. Oleh karenanya perlu dilakukan transformasi pada DataKategori dengan cara mengubah kode numerik dari sebagian atribut, sedemikian sehingga setiap atribut memiliki kode numerik yang berbeda dengan atribut lain. Atribut yang dikenai transformasi adalah Status Kerja, Status Kawin, dan Jenis Kelamin.

Beberapa atribut pada DataNumerik memiliki rentang nilai yang sangat berbeda. Sebagai contoh; atribut SKS memiliki rentang nilai antara 3 sampai dengan 175, sedangkan IPK memiliki rentang nilai antara 0 sampai dengan 4. Hal ini dapat mempengaruhi perhitungan *dissimilarity* antar obyek karena hasil perhitungan akan didominasi oleh perbedaan nilai SKS dibanding IPK (Tan, Steinbach, & Kumar, 2006). Oleh karenanya perlu dilakukan normalisasi terhadap semua atribut DataNumerik untuk mendapatkan nilai yang proporsional tanpa mengubah informasi yang terkandung. Normalisasi yang digunakan adalah *z-score normalization*.

Pengelompokkan DataKategori

Pengelompokkan DataKategori dilakukan dengan 11 (sebelas) variasi nilai *threshold* antara 0,90 dan 1,0. Untuk setiap *threshold* yang dimasukkan, akan menghasilkan jumlah kelompok, anggota tiap-tiap kelompok, dan nilai kohesi untuk masing-masing kelompok. *Threshold* merupakan parameter yang dapat digunakan untuk mengukur *similarity* dari pasangan obyek yang bertetangga. Semakin besar *threshold,* semakin mirip pasangan yang bertetangga tersebut. Hasil pengelompokkan terbaik menghasilkan 44 kelompok, diperoleh pada *threshold* 0,98 dengan nilai kohesi 2044.

Kelompok yang dihasilkan oleh algoritma *QROCK* merupakan *graph based clusters*. Nilai kohesi tiap kelompok didapat dengan cara menghitung jumlah *edge* yang menghubungkan tiap obyek dalam kelompok, dibagi dengan jumlah anggota kelompok. Kualitas hasil pengelompokkan diukur dengan cara menjumlahkan nilai kohesi dari tiap-tiap kelompok yang dihasilkan. Semakin tinggi total nilai kohesi suatu hasil pengelompokkan, semakin baik kelompok yang dihasilkan (Dutta, Mahanta, & Arun., 2005).

Kelompok-kelompok yang dihasilkan, terbentuk berdasarkan kesamaan pada empat atribut, yaitu Pendidikan Akhir, Status Kerja, Status Kawin, dan Jenis Kelamin. Dua obyek akan berada dalam kelompok yang sama apabila memiliki kesamaan pada keempat atribut tersebut. Sebaliknya, apabila salah satu dari keempat atribut tersebut berbeda, maka kedua obyek dipastikan akan berada dalam kelompok yang berbeda. Dua atribut lain yaitu Jurusan Asal dan UPBJJ tidak menentukan terbentuknya kelompok .

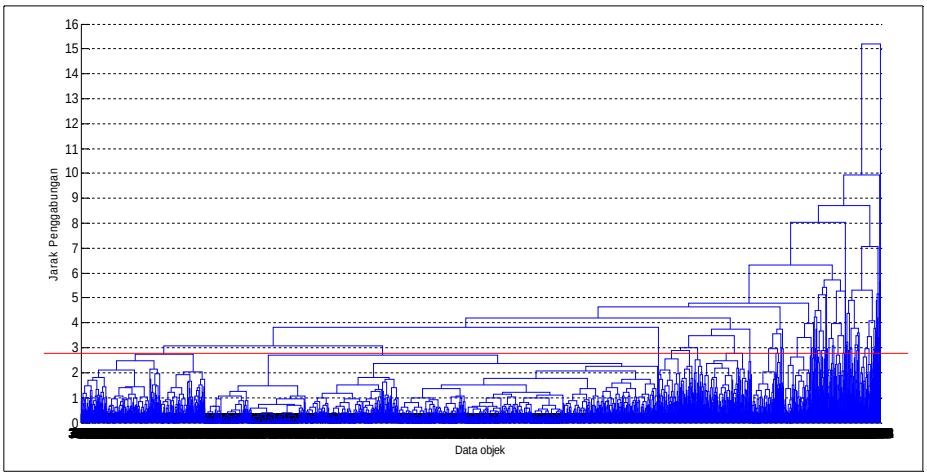
Tabel 1. Karakteristik 6 Kelompok Besar Hasil Pengelompokkan DataKategori

Nomor Kelompok	Jumlah Anggota	Karakteristik Anggota Kelompok
3	1113	Laki-laki bekerja, tidak kawin, SLTA
2	882	Laki-laki tidak bekerja, tidak kawin, SLTA
6	609	Laki-laki tidak bekerja, kawin, SLTA
7	391	Perempuan tidak bekerja, tidak kawin, SLTA
4	365	Perempuan bekerja, tidak kawin, SLTA
11	139	Perempuan bekerja, kawin, SLTA

Dari 44 kelompok yang dihasilkan terdapat 6 kelompok besar (84,7%, dengan anggota kelompok antara 139 sampai dengan 1113 obyek) dan 38 kelompok kecil. Karakteristik 6 kelompok besar tersebut disajikan pada Tabel 1.

Pengelompokkan DataNumerik

Pengelompokkan DataNumerik menggunakan 2 macam ukuran jarak, yaitu *Eucliden distance* dan *Cityblock distance,* dan 3 macam metode penggabungan, yaitu *Single link, Complete link,* dan *Average link*. Terdapat 6 kemungkinan kombinasi ukuran jarak dan metode penggabungan yang digunakan sebagai masukan dari algoritma *AGNES*.



Gambar 3. *Dendrogram* hasil terbaik pengelompokkan DataNumerik

Pengelompokkan terbaik diperoleh dari kombinasi ukuran jarak *Cityblock distance* dan metode penggabungan *Average link* dengan nilai *cophenet* 0,822. Dendrogram hasil pengelompokkan terbaik ditunjukkan oleh Gambar 3. Dengan memotong *dendrogram* pada jarak 2,8 dimana terjadi loncatan jarak penggabungan, diperoleh 69 kelompok.

Dari 69 kelompok yang terbentuk, terdapat 6 kelompok besar (86,7% dari data observasi) dan 63 kelompok kecil. Karakteristik dari 6 kelompok besar disajikan pada Tabel 3.

Tabel 3. Karakteristik 6 Kelompok Besar Hasil Pengelompokkan DataNumerik

Nomor Kelompok	Jumlah Anggota	Karakteristik
18	2340	Usia 23 tahun, 13SKS, IPK 1,52 & Lama Studi 1,8 semester
62	641	Usia 31 tahun, 16SKS, IPK 1,64 & Lama Studi 2,5 semester
28	229	Usia 22 tahun, 44SKS, IPK 2,03 & Lama Studi 5,7 semester
23	136	Usia 42 tahun, 10SKS, IPK 1,40 & Lama Studi 1,9 semester
12	122	Usia 21 tahun, 12SKS, IPK 2,96 & Lama Studi 1,6 semester
47	116	Usia 23 tahun, 29SKS, IPK 1,42 & Lama Studi 7 semester

Secara umum hasil pengelompokkan DataNumerik menunjukkan bahwa lebih dari 78% mahasiswa nonaktif belajar di UT hanya selama dua semester. Tingkat kelulusan matakuliah dalam

dua semester pertama sangat rendah, hal ini ditunjukkan dengan rendahnya SKS dan IPK yang dicapai.

Sebagaimana disampaikan Saxena (2002) bahwa dalam data kemahasiswaan sering tersimpan informasi yang sangat penting tentang mahasiswa. Pada pengelompokkan DataNumerik ditemukan beberapa kelompok kecil yang dapat dikategorikan sebagai *outlier* tetapi perlu mendapat perhatian karena merupakan informasi yang penting bagi pengelola UT. Sebagai contoh, kelompok 4 hanya beranggotakan 1 mahasiswa yang sudah menempuh 145 SKS dengan IPK 2,22 dan telah mengikuti pendidikan di UT selama 12 semester, sedangkan kelompok 16 beranggotakan 2 mahasiswa yang masing-masing telah menempuh 129 SKS dan 132 SKS dengan IPK 2,32 dan 2,11. Jika hanya dilihat dari SKS dan IPK seharusnya mahasiswa dalam kelompok 4 telah memenuhi syarat kelulusan, sedangkan mahasiswa dalam kelompok 16 berpotensi tinggi untuk dapat menyelesaikan studinya. Perlu pemeriksaan lebih lanjut apa yang menyebabkan mahasiswa tersebut berstatus nonaktif.

Pengelompokkan DataGabungan

Hasil pengelompokkan DataKategori menempatkan data obyek ke dalam 44 kelompok, sedangkan hasil pengelompokkan DataNumerik menempatkan data obyek ke dalam 69 kelompok. Struktur masing-masing keluaran tersebut berupa vektor berukuran 4132 (merepresentasikan data obyek) yang berisi nomor kelompok dimana data obyek berada.

DataGabungan dibangun hanya dari enam kelompok besar hasil pengelompokkan DataKategori dan enam kelompok besar hasil pengelompokkan DataNumerik. Alasannya bahwa kelompok-kelompok tersebut merepresentasikan keluaran dari masing-masing proses pengelompokkan karena mewakili lebih dari 80% keseluruhan data observasi.

Anggota dari DataGabungan adalah irisan enam kelompok besar hasil pengelompokkan DataKategori dan enam kelompok besar hasil pengelompokkan DataNumerik. DataGabungan berbentuk matrik 3069x2 dengan atribut pertama berisi nomor kelompok dari hasil pengelompokkan DataKategori dan atribut kedua berisi nomor kelompok dari hasil pengelompokkan DataNumerik.

Kelompok yang dihasilkan oleh suatu algoritma pengelompokkan menempatkan setiap data obyek ke dalam satu kelompok tertentu. Jika dua obyek berada dalam kelompok yang sama maka kedua obyek tersebut dianggap sama. Sebaliknya jika dua obyek berada dalam kelompok yang berbeda maka kedua obyek dianggap berbeda. Jelas bahwa kelompok yang dihasilkan oleh setiap algoritma pengelompokkan tidak dapat diurutkan sebagaimana mengurutkan bilangan riil. Oleh karenanya kelompok-kelompok tersebut dapat dipandang sebagai data kategori. Zengyou *et al.* (2002) menyampaikan bahwa karena keluaran dari masing-masing algoritma klastering merupakan data kategori, maka persoalan *cluster ensemble* dapat dipandang sebagai persoalan pengelompokkan data kategori. Hasil dari masing-masing algoritma pengelompokkan dapat digabung menjadi data baru dengan tipe kategori. Karena itulah pengelompokkan DataGabungan menggunakan algoritma *QROCK*.

Pengelompokkan DataGabungan dilakukan dengan lima variasi nilai *threshold* 0,6, 0,7, 0,8, 0,9 dan 1,0. Hasil terbaik pengelompokkan diperoleh pada *threshold* 1,0, menghasilkan 35 kelompok. Dari 35 kelompok yang dihasilkan, terdapat 7 kelompok besar (78% dari data observasi) dan 28 kelompok kecil. Karakteristik 7 klaster besar tersebut, sebagaimana disajikan pada Tabel 4, merupakan kombinasi dari lima klaster terbesar hasil klastering DataKategori dan dua klaster terbesar hasil klastering DataNumerik.

Dapat dikatakan bahwa Tabel 4 merepresentasikan karakteristik kelompok-kelompok mahasiswa nonaktif Program Studi Matematika tahun 2008. Kelompok-kelompok yang terbentuk memiliki kesamaan pada pendidikan akhir (SLTA), status pekerjaan, status perkawinan, dan jenis kelamin. Faktor prestasi akademik menunjukkan bahwa dua semester pertama merupakan masa kritis bagi mahasiswa Program Studi Matematika UT. Tingkat kelulusan matakuliah dalam dua semester pertama sangat rendah. Hal ini ditunjukkan dengan rendahnya SKS dan IPK yang dicapai.

Tabel 4. Karakteristik 7 Kelompok Terbesar Hasil Pengelompokkan DataGabungan

Nomor Kelompok	Jumlah Anggota	Karakteristik Anggota Kelompok
2	745	Laki-laki bekerja, tidak kawin, 23th, 13SKS,IPK1,52& 1,8 semester
1	617	Laki-laki tidak bekerja, tidak kawin, 23th, 13SKS,IPK1,52& 1,8 semester
6	278	Perempuan tdk bekerja, tdk kawin, 23th, 13SKS,IPK1,52& 1,8 semester
5	252	Perempuan bekerja, tidak kawin, 23th, 13SKS,IPK1,52& 1,8 semester
11	234	Laki-laki tidak bekerja, kawin, 31th, 16SKS,IPK1,64 & 2,5 semester
4	158	Laki-laki tidak bekerja, kawin, 23th, 13SKS,IPK1,52& 1,8 semester
9	120	Laki-laki bekerja, tidak kawin, 31th, 16SKS,IPK1,64 & 2,5 semester

Hasil pengelompokkan ini semestinya dapat digunakan oleh pengelola untuk membuat program penanganan terhadap mahasiswa nonaktif lebih tepat. Karakteristik kelompok tertentu yang dihasilkan dapat menjadi acuan awal dalam merangcang program penanganan pelayanan mahasiswa. Sedangkan jumlah anggota kelompok dapat dijadikan ukuran besaran program pelayanan tersebut.

Sebagaimana telah disebutkan sebelumnya bahwa jumlah kelompok hasil pengelompokkan DataNumerik diperoleh dengan cara memotong *dendrogram* pada jarak dimana terjadi loncatan tertinggi pada diagram batang daun. Untuk data dengan ukuran kecil tidaklah menjadi soal, tetapi untuk data dengan ukuran besar sebagaimana yang digunakan dalam penelitian ini, hal tersebut menjadi kendala tersendiri. Perlu dipertimbangkan penggunaan ukuran kuantitatif untuk menentukan jarak pemotongan tersebut, misalkan nilai inkonsistensi dari setiap *link* pada pohon yang dihasilkan oleh algoritma *AGNES.*

Kompleksitas dari *algCEBMDC* dipengaruhi oleh tiga komponen, yaitu kompleksitas pengelompokkan data kategori, kompleksitas pengelompokkan data numerik, dan kompleksitas pengelompokkan data gabungan. Dengan kata lain kompleksitas dari *algCEBMDC* ditentukan oleh kompleksitas dari algoritma yang digunakan pada masing-masing komponen (Zengyou *et al.*, 2002). Dengan menggunakan kombinasi algoritma *QROCK* dan Algoritma *AGNES*, kompleksitas algoritma *algCEBMDC* pada penelitian ini adalah $O(n^3)$.

SIMPULAN

Algoritma *algCEBMDC* yang digunakan dalam penelitian ini menerapkan algoritma *QROCK* untuk pengelompokkan data kategori dan algoritma *AGNES* untuk pengelompokkan data numerik. Kelompok-kelompok yang terbentuk memiliki kesamaan pada pendidikan akhir (SLTA), status pekerjaan, status perkawinan, dan jenis kelamin. Faktor prestasi akademik menunjukkan bahwa dua semester pertama merupakan masa kritis bagi mahasiswa Program Studi Matematika UT. Tingkat kelulusan matakuliah dalam dua semester pertama sangat rendah.

REFERENSI

Chong, H.Y., Samuel D., Angel J.P., & Charles K. (2010). A data mining approach fo identifying predictor of student retention from sophomore to junior year. *Journal of Data Science*. 8, 307-325.

Dutta, M., Mahanta A.K., & Arun K.P. (2005). QROCK: A quick version of the ROCK algorithm for clustering of categorical data. *Proceedings of the15IEEE International Conference on Data Engineering*, 2004.

Han, J., & Kamber M. (2001). Data mining: Concepts and techniques. USA: *Academic Press*.

Marisa, A. (2008). *Perbandingan algoritme clustering rock dan qrock untuk data kategorik.* Skripsi sarjana yang tidak dipublikasikan. Institut Pertanian Bogor, Bogor:

Saxena, A., Pankaj K., & Suresh G. (2002). *Aplication of cluster analysis as a tool to analyse distance educations students*. Indira Gandi Open University, New Delhi, India.

Shaeela, A., Tasleem M., & Ahsan R.S. (2010). Data mining model for higher education system. *Europen Journal of Scientific Research,* 43(1), 24-29.

Tan, P., Steinbach M., & Kumar V. (2006). *Introduction to data mining*. USA: Pearson Education,Inc

Zengyou, H., Xiaofe I X., & Shengchum D. (2002). *Clustering mixed numeric and categorical data: A cluster Ensemble Approach.* http://arxiv.org/ftp/cs/papers/0509/050911.pdf