CLUSTERING MENGGUNAKAN *SELF ORGANIZING MAPS* (STUDI KASUS: DATA PPMB IPB)

Irman Hermadi [1], Imas S. Sitanggang [1], Edward [2]
[1] Staf Departemen Ilmu Komputer, Fakultas Matematika dan IPA, Institut Pertanian Bogor
[2] Mahasiswa Departemen Ilmu Komputer, Fakultas Matematika dan IPA, Institut Pertanian Bogor

ABSTRAK

Data berukuran besar yang sudah disimpan tidak digunakan secara optimal karena manusia seringkali tidak memiliki waktu dan ilmu yang cukup untuk mengelolanya. Kasus ini terjadi di Panitia Penerimaan Mahasiswa Baru Institut Pertanian Bogor (PPMB IPB). Penelitian ini bertujuan untuk mengimplementasikan Algoritma Self Organizing Maps (SOM) dalam clustering data, dan untuk mendapatkan karakteristik data dari hasil clustering.

Data yang digunakan adalah rata-rata nilai Biologi, Fisika, Matematika, dan Kimia (cawu 1 sampai cawu 7) dari pelamar tahun 2004 dengan pilihan pertama program sarjana di Fakultas Pertanian, IPB. Data (sebanyak 1899 baris dan 4 field yaitu: Biologi, Fisika, Kimia, dan Matematika) akan menjadi masukan algoritma SOM, dengan parameter awal algoritma SOM: ukuran vektor bobot/ output: 3, 4, 5, 6, 7, 8, 9, 10; learning rate: 0.1, 0.5, 0.9; ukuran lingkungan: 0, dan penurunan learning rate: 0.1, 0.5, 0.9, 1. Penentuan bobot pemenang dalam algoritma SOM menggunakan Jarak Mahalanobis, dengan fungsi topologi adalah Gridtop, dan inisialisasi nilai bobot awal dengan nilai midpoint. Kriteria pemberhentian algoritma SOM dalam penelitian ini adalah iterasi, dengan banyak iterasi: 1, 5, dan 10. Hasil clustering dari SOM divalidasi menggunakan Indeks Davies-Bouldin.

Hasil clustering data yang memiliki DBI minimal (53.472) dari penelitian adalah ukuran vektor bobot 9 dengan learning rate 0.9, penurunan learning rate 0.1, dan 5 iterasi. Pelamar dari Sumatera banyak berada pada cluster yang memiliki rataan nilai Biologi, Fisika, Kimia, dan Matematika lebih tinggi (81.12, 77.50, dan 74.16). Berbeda dengan daerah asal Jawa, yang banyak berada di cluster yang memiliki rataan lebih rendah (74.08, 73.09, 71.91, 70.04, 68.59, dan 67.93). Pelamar dari Luar Negeri tergolong pelamar dengan nilai rendah, hanya berada di cluster dengan rataan 68.59.

Peluang diterima dari masing-masing kategori SMA bergantung kepada nilai, namun nilai pelamar bukan satu-satunya acuan dalam seleksi penerimaan mahasiswa baru. Kategori SMA juga berkontribusi terhadap diterima/tidaknya pelamar. Penelitian selanjutnya dapat difokuskan untuk optimasi kombinasi nilai-nilai parameter algoritma SOM.

Kata kunci: *Self Organizing Maps*, Jarak *Mahalanobis, Indeks Davies Bouldin*, Analisis *Cluster*.

1. PENDAHULUAN

Latar Belakang

Perkembangan teknologi telah mengakibatkan meningkatnya data dalam jumlah besar. Data berukuran besar yang sudah disimpan tidak digunakan secara optimal karena manusia seringkali tidak punya waktu dan ilmu yang cukup untuk mengelolanya.

Kasus ini terjadi di Panitia Penerimaan Mahasiswa Baru Institut Pertanian Bogor (PPMB IPB). PPMB IPB mengumpulkan data pelamar program sarjana setiap tahun, meliputi data akademik, data penilaian terhadap sekolah asal, serta data pribadi. Data pelamar disimpan setelah digunakan untuk menyeleksi calon mahasiswa baru IPB.

Data mining sangat sesuai untuk diterapkan pada data berukuran besar. Penerapan *data mining* pada data PPMB IPB diharapkan bisa menambang ilmu pengetahuan dan informasi yang penting dan berguna untuk pengambilan keputusan di masa depan.

Metode *data mining* yang akan diterapkan dalam penelitian ini adalah *clustering* dengan menggunakan algoritma *Self Organizing Maps* (SOM). *Clustering* digunakan untuk melakukan pengelompokan data tanpa berdasarkan target variabel kelas data tertentu. Bahkan *clustering* dapat dipakai untuk memberikan label pada kelas data yang belum diketahui. *Clustering* menggunakan *Fuzzy C-Means* pernah dilakukan pada data pelamar melalui jalur Undangan Seleksi Masuk IPB (USMI) yang terpilih di Fakultas Pertanian (Wisnujati 2006). Hasil *clustering* yang lebih baik dari penelitian sebelumnya diharapkan bisa didapatkan dan bisa memberikan ilmu pengetahuan dan informasi yang berguna.

Tujuan Penelitian

Penelitian ini memiliki tujuan:

1 mengimplementasikan algoritma SOM dalam c*lustering* data pelamar jalur USMI tahun 2004 dengan pilihan pertama program studi di Fakultas Pertanian IPB,
2 mendapatkan karakteristik data dari hasil *clustering* menggunakan SOM.

Ruang Lingkup

Penelitian ini meliputi penerapan salah satu fungsionalitas dari *data mining* yaitu analisis *cluster*. Analisis *cluster* menggunakan metode SOM akan diimplementasikan pada rata-rata nilai Biologi, Fisika, Matematika, dan Kimia (cawu 1 sampai cawu 7) dari pelamar jalur USMI tahun 2004 dengan pilihan pertama program sarjana di Fakultas Pertanian, IPB. Persentase masing-masing *cluster* berdasarkan daerah asal pelamar, kategori SMA, dan putusan diterima akan dilihat untuk melihat pola yang mungkin terjadi dari *clustering*.

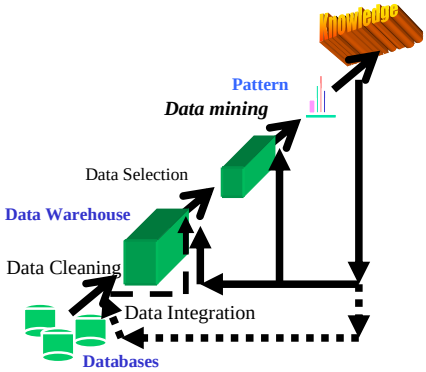
Manfaat Penelitian

Informasi yang bernilai berupa karakteristik pelamar dengan pilihan pertama program studi di Fakultas Pertanian IPB akan dihasilkan dari penelitian ini. Analisis *cluster* menggunakan SOM diharapkan akan bermanfaat sebagai pertimbangan pengambilan keputusan di masa depan.

2. TINJAUAN PUSTAKA

Knowledge Discovery in Database

Data mining merupakan salah satu tahap pada proses *Knowledge Discovery in Database* (KDD). KDD adalah penyulingan informasi menarik yang tidak biasa, yang terkandung dalam basis data berukuran besar, yang sebelumnya tidak diketahui dan potensial bermanfaat (Han & Kamber 2001).



Gambar 1 Tahapan dalam KDD (Han & Kamber 2001).

Tahap-tahap proses KDD (Gambar 1) menurut (Han & Kamber 2001), adalah:

1 Pembersihan Data
2 Integrasi Data
3 Seleksi Data
4 Transformasi Data
5 *Data mining*
6 Evaluasi Pola
7 Presentasi Pengetahuan

Pembersihan Data

Data yang bersih adalah data yang konsisten dan tidak mengandung nilai yang tidak lengkap dan *noise*. Proses pembersihan data bertujuan untuk melengkapi nilai yang tidak lengkap, memperhalus *noise* ketika teridentifikasi, dan memperbaiki ketidakkonsistenan data. Secara umum data yang tidak bersih adalah: nilai yang tidak lengkap, data yang mengandung *noise*, dan data yang tidak konsisten (Han & Kamber 2001).

Integrasi dan Transformasi Data

Integrasi data mengkombinasikan data dari sumber-sumber yang berbeda menjadi bentuk sebuah penyimpanan data yang koheren, seperti dalam data *warehousing*. Proses transformasi data mengubah data menjadi bentuk yang sesuai untuk dilakukan tahapan *data mining.* Proses ini meliputi: penghalusan, agregasi, generalisasi dari data, normalisasi, dan konstruksi atribut (atau konstruksi fitur) (Han & Kamber 2001).

Data Mining

Data mining adalah kegiatan penemuan pola-pola yang menarik dari data berukuran besar yang disimpan dalam basis data, data *warehouse*, atau sarana penyimpanan yang lain. *Data mining* dapat diklasifikasikan menjadi dua kategori: *descriptive data mining* dan *predictive data mining*. *Descriptive data mining* menjelaskan himpunan data dengan memberikan banyak informasi secara jelas dalam kalimat yang singkat dan memberikan sifat-sifat umum yang menarik dari data. *Predictive data mining* menganalisis data yang bertujuan untuk membangun sebuah atau himpunan model, dan berusaha untuk meramalkan karakteristik dari himpunan data baru (Han & Kamber 2001).

Menurut (Han & Kamber 2001), fungsionalitas *data mining* adalah:

1 Deskripsi kelas/ deskripsi konsep dan diskriminasi,
2 Analisis asosiasi,
3 Klasifikasi dan prediksi,
4 Analisis *cluster*,

5 Analisis pencilan, dan
6 Analisis evolusi.

Analisis *Cluster*

Clustering adalah pengelompokan dari *record*, observasi-observasi atau kasus-kasus ke kelas yang memiliki kemiripan objek-objeknya. *Cluster* adalah koleksi dari *record* yang mirip, dan tidak mirip dengan *record* dari *cluster* lain. *Clustering* berbeda dengan klasifikasi, dalam hal tidak ada variabel target untuk *clustering*. *Clustering* tidak mengklasifikasikan, meramalkan, atau memprediksi nilai dari sebuah variabel target. Algoritma-algoritma *clustering* digunakan untuk menentukan segmen keseluruhan himpunan data menjadi *subgroup* yang relatif sama atau *cluster*, dengan kesamaan *record* dalam *cluster* dimaksimumkan dan kesamaan *record* di luar *cluster* diminimumkan (Larose 2005).

Secara umum metode utama *clustering* dapat diklasifikasikan menjadi kategori-kategori berikut (Han & Kamber 2001):

- Metode partisi. Misalkan ada sebuah basis data berisi n objek. Metode partisi membangun k partisi pada basis data tersebut, dengan tiap partisi merepresentasikan *cluster* dan k ≤ n. Partisi yang terbentuk harus memenuhi syarat *yaitu* setiap *cluster* harus berisi minimal satu objek dan setiap objek harus termasuk tepat satu *cluster*.
- Metode hirarkhi, yaitu membuat sebuah dekomposisi berhirarki dari himpunan data (atau objek) menggunakan beberapa kriteria. Metode ini memiliki dua jenis pendekatan yaitu :
  - *Agglomerative*, dimulai dengan titik-titik sebagai *cluster* individu. Pada setiap tahap dilakukan penggabungan setiap pasangan titik pada *cluster* sampai hanya satu titik (atau *cluster*) yang tertinggal.
  - *Divisive*, dimulai dengan satu *cluster* besar yang berisi semua titik data. Pada setiap langkah, dilakukan pemecahan sebuah *cluster* sampai setiap *cluster* berisi sebuah titik (atau terdapat k *cluster*).
- Metode berdasarkan kepekatan, merupakan pendekatan yang berdasarkan pada konektivitas dan fungsi kepadatan.
- Metode berdasarkan *grid*, merupakan pendekatan yang berdasarkan pada struktur *multiple-level granularity*.
- Metode berdasarkan model, yaitu: sebuah model yang dihipotesis untuk tiap *cluster* dan ide dasarnya adalah untuk menemukan model yang cocok untuk tiap *cluster*.

***Self Organizing Maps* (SOM)**

Jaringan Kohonen diperkenalkan oleh Teuvo Kohonen seorang ilmuwan Finlandia pada tahun 1982. Jaringan Kohonen memberikan sebuah tipe dari SOM; kelas khusus dari jaringan syaraf tiruan (Larose 2004). SOM merupakan metode berdasarkan model dari pendekatan jaringan syaraf tiruan (Han & Kamber 2001).

SOM adalah metode terkemuka pendekatan jaringan syaraf tiruan untuk *clustering*, setelah *competitive learning* (Han & Kamber 2001). SOM berbeda dengan *competitive learning* yaitu syaraf dalam satu lingkungan belajar untuk mengenali bagian lingkungan dari ruang input. SOM mengenali distribusi (seperti *competitive learning*) dan topologi dari vektor input yang melalui proses *training* (Demuth & Beale 2003). SOM memperlihatkan tiga karakteristik: kompetisi yaitu setiap vektor bobot saling berlomba untuk menjadi simpul pemenang, kooperasi yaitu setiap simpul pemenang bekerjasama dengan lingkungannya, dan adaptasi yaitu perubahan simpul pemenang dan lingkungannya (Larose 2004).

Algoritma *Self Organizing Maps*

Misalkan himpunan dari m nilai-nilai *field* untuk *record* ke-n menjadi sebuah vektor input $x_n = x_{n1}, x_{n2}, x_{n3}, \cdots, x_{nm}$, dan himpunan dari m bobot untuk simpul *output* tertentu j menjadi vektor bobot $w_j = w_{1j}, w_{2j}, \ldots, w_{mj}$ (Larose 2004).

Berikut ini adalah langkah-langkah algoritma SOM (Larose 2004):

Untuk setiap vektor x, lakukan:

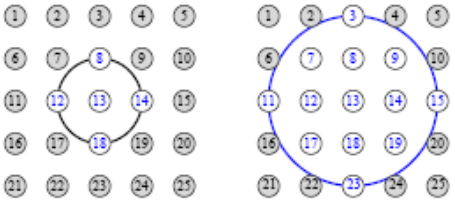
- Kompetisi. Untuk setiap simpul output j, hitung nilai $D(w_j, x_n)$ dari fungsi jarak. Tentukan simpul pemenang J yang meminimumkan $D(w_j, x_n)$ dari semua simpul *output*.
- Kooperasi. Identifikasikan semua simpul *output* j dalam lingkungan simpul pemenang J didefinisikan oleh lingkungan berukuran R. Untuk simpul-simpul ini, lakukan:
  - Adaptasi. Perbaharui nilai bobot: $w_{ij,new} = w_{ij,current} + \eta\left(x_{ni} - w_{ij,current}\right)$.
- Perbaharui *learning rate* (η) dan ukuran lingkungan (R) seperlunya.
- Hentikan perlakuan ketika kriteria pemberhentian dicapai.

Keterangan:

- Inisialisasi nilai bobot biasanya menggunakan nilai tengah (*middle*

point/midpoint) atau menggunakan nilai acak (Demuth & Beale 2003).

- Lingkungan berukuran R berisi indeks dari semua simpul-simpul yang berada dalam radius R dari simpul pemenang i^*. $N_i(d) = \{j, d_{ij} \le R\}$ (Demuth & Beale 2003).



Gambar 2 Ilustrasi lingkungan (Demuth & Beale 2003).

Gambar 2 mengilustrasikan konsep lingkungan. Gambar 2 kiri menunjukkan lingkungan dari radius R=1 sekeliling simpul 13. Gambar 2 kanan menunjukkan lingkungan dari radius R=2. Topologi lingkungan yang umum digunakan ada 3: topologi grid, topologi hexagonal, dan topologi random (Demuth & Beale 2003).

- Fungsi jarak biasanya digunakan Jarak Euclidean

$$D(w_j, x_n) = \sqrt{\sum_i (w_{ij} - x_{ni})^2}$$

(Kaski 1997, Demuth & Beale 2003).

- Jarak Mahalanobis digunakan untuk atribut yang berkorelasi satu sama lain

$$D(w_j, x_n) = (w_j - x_i)\Sigma^{-1}(w_j - x_i)^T,$$

dengan Σ ialah matriks kovarian dari vektor input (x_n),

$$\Sigma_{j,k} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \overline{X_j})(X_{ik} - \overline{X_k})$$

(Tan *et al*. 2004).

- Perubahan tingkat pembelajaran (LR/ α / η) $0 < \eta < 1$, dengan rumus $\alpha(t+1) = \theta\,\alpha(t)$. Lambang θ adalah penurunan tingkat pembelajaran (PLR), menurun seiring perubahan waktu *t* (Laurence 1994).
- Kriteria pemberhentian bisa berupa pembatasan jumlah iterasi, atau ketika $\eta = 0$ (Larose 2004).

Validitas *Cluster*

Validasi *cluster* ialah prosedur yang mengevaluasi hasil analisis cluster secara kuantitatif dan objektif (Jain & Dubes 1988). Terdapat tiga pendekatan untuk mengeksplorasi validitas *cluster*:

1 kriteria eksternal, mengevaluasi hasil dari metode *clustering* berdasarkan pra-spesifikasi struktur yang diterima dari sebuah data yang mencerminkan intuisi pengguna tentang struktur *clustering* dari data,
2 kriteria internal, mengevaluasi hasil *clustering* dalam konsep kuantitatif yang didapat dari data, dan
3 kriteria relatif, membandingkan sebuah struktur *clustering* dengan struktur *clustering* yang lain yang didapatkan dari metode *clustering* yang sama tetapi nilai-nilai parameternya dimodifikasi (Salazar *et al*. 2002).

Untuk memilih skema *clustering* optimal, ada dua kriteria (Salazar *et al*. 2002):

1 *Compactness*, yaitu anggota dari masing-masing *cluster* harus sedekat mungkin dengan yang lain, dan
2 *Separation*, yaitu *cluster* harus terpisah secara luas dari *cluster* lain.

Indeks validitas digunakan sebagai metode validasi *cluster* untuk evaluasi kuantitatif dari hasil *clustering* (Salazar *et al*. 2002). Beberapa indeks yang biasa digunakan adalah: Hubert Statistic, Indeks Dun, Indeks Davies-Bouldin, Root-mean-square *standard deviation* (RMSSTD), dan R-squared (RS) (Salazar *et al*. 2002).

Indeks Davies-Bouldin

Pendekatan pengukuran ini untuk memaksimalkan jarak *inter-cluster* di antara *Cluster* C_i dan C_j dan pada waktu yang sama mencoba untuk meminimalkan jarak antara titik dalam sebuah *cluster*. Jarak *intra-cluster* $s_c(Q_k)$ dalam *Cluster* Q_k ialah

$$s_c(Q_k) = \frac{\sum_i \|X_i - C_k\|}{N_k},$$

dengan N_k adalah banyak titik yang termasuk dalam *Cluster* Q_k dan C_k adalah *centroid* dari *Cluster* Q_k. Jarak *Inter-cluster* didefinisikan sebagai

$$d_{kl} = \|C_k - C_l\|,$$

dengan C_k dan C_l ialah *centroid Cluster* ***k*** dan *Cluster* ***l***. Dilain pihak, Indeks Davies-Bouldin didefinisikan sebagai

$$DB(nc) = \frac{1}{n}\sum_{\substack{k=1 \\ k \ne l}}^{nc} \max\left\{\frac{s_c(Q_k) + s_c(Q_l)}{d_{kl}(Q_k, Q_l)}\right\},$$

dengan *nc* ialah banyak *cluster*. Skema *clustering* yang optimal menurut Indeks Davies-

Bouldin adalah yang memiliki Indeks Davies-Bouldin minimal (Salazar *et al*. 2002).

3. METODE PENELITIAN

Praproses

Penelitian ini akan dilakukan menggunakan proses KDD. Tahapan yang termasuk dalam praproses yaitu: pembersihan data, integrasi data, transformasi data, dan seleksi data. Tahap pembersihan data, integrasi data, dan transformasi data telah dilakukan oleh peneliti sebelumnya (Riyanti 2005).

Pada tahap Seleksi data akan dipilih rata-rata nilai Biologi, Fisika, Matematika, dan Kimia (cawu 1 sampai cawu 7) dari pelamar tahun 2004 dengan pilihan pertama program sarjana di Fakultas Pertanian, IPB. Pemilihan atribut nilai Biologi, Fisika, Kimia, dan Matematika karena atribut ini selalu diacu dalam seleksi penerimaan mahasiswa baru jalur USMI (Wisnujati 2006). Data hasil seleksi sebanyak 1899 baris dan 4 *field* yaitu: Biologi, Fisika, Kimia, dan Matematika.

Data Mining

Data mining yang dilakukan pada penelitian ini adalah *clustering* data menggunakan algoritma SOM untuk melihat karakteristik (deskripsi) pelamar tahun 2004 dengan pilihan pertama program studi yang ada di Fakultas Pertanian IPB melalui jalur USMI. Tahapan data mining dilakukan menggunakan aplikasi yang dibangun.

Masukan ke algoritma SOM adalah data dari praproses dengan kombinasi dari parameter awal. Parameter awal dari algoritma SOM yang akan digunakan adalah:

1 ukuran (j) dari vektor bobot (w_j) : 3, 4, 5, 6, 7, 8, 9, dan 10,
2 *learning rate* (η) : 0.1, 0.5, dan 0.9,
3 ukuran lingkungan (R) : 0, dan
4 penurunan *learning rate* (θ) : 0.1, 0.5, 0.9, dan 1.

Metode inisialisasi nilai vektor bobot menggunakan *midpoint* dengan topologi yang digunakan adalah topologi grid. Jarak Mahalanobis digunakan sebagai fungsi jarak karena antar atribut saling berkorelasi (Wisnujati 2006). Kriteria pemberhentian algoritma SOM dalam penelitian ini adalah iterasi, dengan banyak iterasi: 1, 5, dan 10.

Seluruh hasil *clustering* dari algoritma SOM akan divalidasi menggunakan validasi *cluster* Indeks Davies-Bouldin (DBI). Dari berbagai kombinasi parameter awal dan iterasi, akan dipilih *clustering* yang menghasilkan DBI minimal sebagai *clustering* terbaik.

Deskripsi Aplikasi *Self Organizing Maps*

Aplikasi *Self Organizing Maps* dibangun untuk digunakan pada tahap data mining. Aplikasi ini memiliki menu:

- *Home*,
  - *Resume*,
  - Tabel Input,
  - *Scatter Graph*,
  - Frekuensi *Graph*,
  - *Centroid* dan Rataan,
  - Bobot *Graph*,
  - *Begin New Train*, dan
  - *Reset* Eksekusi,
- Arsip,
  - *Resume*,
  - Tabel Input,
  - *Scatter Graph*,
  - Frekuensi *Graph*,
  - *Centroid* dan Rataan,
  - Bobot *Graph*,
  - *Begin New Train*, dan
  - *Reset* Eksekusi,
- PHP,
- *Help*, dan
- *About*.

Menu *Home* digunakan untuk menampilkan form isian parameter awal dan data yang akan digunakan. Menu *Home* memiliki anak menu yang sama dengan menu Arsip. Menu Arsip digunakan untuk menampilkan parameter awal, data, DBI, dan waktu dari perlakuan yang pernah dilakukan. Menu *Resume* untuk menampilkan parameter awal, data, DBI, dan waktu. Menu Tabel Input untuk menampilkan data yang telah di lakukan tahap *clustering* dengan parameter awal, DBI, dan waktu di Menu *Resume*. Menu *Scatter Graph* untuk menampilkan data dan *centroid* dalam bentuk *scatter plot*. Menu Frekuensi *Graph* untuk menampilkan banyaknya data untuk masing-masing *cluster* dalam bentuk grafik batang. Menu *Centroid* dan Rataan digunakan untuk menampilkan tabel *centroid* dan rataan dari hasil *clustering*. Menu Bobot *Graph* digunakan untuk menampilkan *graph* dari bobot/*centroid* masing-masing *cluster* dalam bentuk grafik titik garis.

Menu PHP digunakan sebagai penghubung ke halaman *phpMyAdmin*. Menu *About* digunakan untuk menampilkan halaman tentang aplikasi secara singkat.

Representasi Pengetahuan

Representasi pengetahuan akan dilakukan terhadap *cluster* yang sudah divalidasi. Representasi pengetahuan akan memperlihatkan karakteristik *cluster* dari SOM berupa rataan dan *centroid* dari *cluster*. Persentase masing-masing *cluster* berdasarkan daerah asal pelamar, kategori SMA, dan putusan diterima akan dilihat untuk melihat pola yang mungkin terjadi dari hasil *clustering*.

Lingkungan Penelitian

Lingkungan penelitian yang digunakan adalah sebagai berikut:

- Perangkat lunak: Microsoft® Windows XP Professional 2002 SP2, Microsoft® Internet Explorer 6.0, PHP 5.0.3, Apache Webserver.
- Perangkat keras: komputer personal dengan spesifikasi Pentium IV 2.4 GHz, RAM 512 MB.

4. HASIL DAN PEMBAHASAN

Indeks Davies Bouldin (DBI)

Pengamatan terhadap DBI dilakukan untuk mengukur validitas dari hasil *clustering*. Parameter penurunan *learning rate* (PLR) dari learning rate (LR) akan berpengaruh terhadap DBI mulai pada iterasi 2. Hal ini bisa dilihat dengan PLR yang berbeda pada iterasi 1 akan menghasilkan DBI yang sama. DBI terbaik untuk masing-masing ukuran *output*/ vektor bobot dapat dilihat pada Tabel 1.

Tabel 1 Indeks Davies-Bouldin terbaik untuk tiap ukuran *output*.

Ukuran Output	LR	PLR	ITERASI	DBI
3	0.9	-	1	209.285
4	0.1	0.5	5	353.452
5	0.5	-	1	202.856
6	0.5	0.1	5	164.302
7	0.5	0.9	10	113.370
8	0.1	-	1	87.917
9	0.9	0.1	5	53.472
10	0.1	0.9	5	79.743

DBI Terbaik

Dari hasil penelitian, Indeks Davies-Bouldin terbaik dihasilkan dengan parameter awal: ukuran *output* 9, LR 0.9, PLR 0.1, dan 5 iterasi, yang menghasilkan DBI 53.472 (Tabel 1). Banyaknya data masing-masing *cluster* dengan ukuran *output* 9 dapat dilihat pada Tabel 2 (penomoran *cluster* tidak menunjukkan tingkatan). Rataan dan *centroid* masing-masing *cluster* dengan ukuran *output* 9 dapat dilihat pada Tabel 3 dan Tabel 4.

Tabel 2 Banyak anggota masing-masing *cluster* dengan ukuran *output* 9.

Cluster ke-	Banyak anggota	Persentase banyak anggota
1	188	9.90
2	284	14.96
3	197	10.37
4	212	11.16
5	199	10.48
6	243	12.80
7	272	14.32
8	189	9.95
9	115	6.06

Tabel 3 *Centroid* masing-masing *cluster* dengan ukuran *output* 9.

Cluster ke-	*Centroid*			
	Biologi	Fisika	Kimia	Matematika
1	69.62	69.76	65.20	66.34
2	68.49	67.43	68.89	69.51
3	83.39	80.48	81.34	79.50
4	74.15	76.03	74.74	71.52
5	74.77	68.84	76.28	76.33
6	77.12	70.97	70.43	75.55
7	75.84	67.27	69.41	66.65
8	73.38	68.79	76.81	68.22
9	73.18	77.11	79.35	82.17

Tabel 4 Rataan nilai mata ajaran masing-masing *cluster* dengan ukuran *output* 9.

Cluster ke-	Rataan				
	Biologi	Fisika	Kimia	Matematika	Rataan
3	83.31	80.37	81.13	79.66	81.12
9	73.33	76.86	78.61	81.20	77.50
5	74.99	68.83	76.57	76.26	74.16
4	74.11	75.87	74.61	71.73	74.08
6	76.28	70.74	69.92	75.41	73.09
8	73.47	68.82	76.94	68.42	71.91
7	76.12	67.44	69.68	66.94	70.04
2	68.60	67.39	68.88	69.50	68.59
1	69.78	69.88	65.46	66.60	67.93
Rataan	74.44	71.80	73.53	72.86	73.16

Deskripsi *Clustering* Terbaik

Cluster 3 yang memiliki 10.37% dari data (Tabel 2), adalah *cluster* yang memiliki rataan Biologi, Fisika, Kimia, dan Matematika tertinggi (Tabel 4). Namun *Cluster* 3 bukan *cluster* yang memiliki nilai yang terbaik untuk seluruh atribut, peringkat ke dua untuk nilai Matematika (Tabel 5).

Tabel 5 Urutan *cluster* berdasarkan nilai

Peringkat	*Cluster* ke-			
	Biologi	Fisika	Kimia	Matematika
1	3	3	3	9
2	6	9	9	3
3	7	4	8	5
4	5	6	5	6
5	4	1	4	4
6	8	5	6	2
7	9	8	7	8
8	1	7	2	7
9	2	2	1	1

Cluster 9 yang memiliki 6.06% dari data (Tabel 2), menduduki peringkat ke dua dari rataan secara keseluruhan (Tabel 4). *Cluster* 9 memiliki nilai Matematika tertinggi, namun hanya menduduki peringkat ke dua dari nilai Fisika dan Kimia, bahkan ke tujuh untuk nilai Biologi (Tabel 5). *Cluster* 9 memiliki kemampuan yang cukup kuat untuk nilai Matematika, Fisika, dan Kimia, namun lemah di Biologi.

Cluster 5 (10.48% dari data) adalah *cluster* yang menduduki peringkat ke tiga dari rataan (Tabel 2 dan Tabel 4). *Cluster* 5 menduduki peringkat ke tiga untuk nilai Matematika, peringkat ke 4 untuk nilai Biologi dan Kimia, dan peringkat ke 6 untuk nilai Fisika (Tabel 5). *Cluster* 5 memiliki kelemahan di nilai Fisika. Nilai Fisika *Cluster* 5 di bawah rata-rata, yaitu 68.83% dari rata-rata 71.80% (Tabel 4).

Cluster 4 (11.16% dari data) adalah *Cluster* yang menduduki peringkat ke empat dari rataan keseluruhan (Tabel 2 dan Tabel 4). *Cluster* 4 memiliki kelebihan di nilai Fisika (menduduki peringkat ke 3), sedangkan untuk nilai Biologi, Kimia, dan Matematika, *Cluster* 4 menduduki peringkat ke lima (Tabel 5).

Cluster 6 menempati peringkat ke 5 untuk rataan keseluruhan (Tabel 4), memiliki anggota terbanyak ke 2 dari data yaitu 12% (Tabel 2). *Cluster* 6 memiliki kemampuan lebih di bidang Biologi dengan peringkat ke dua untuk nilai Biologi (Tabel 5). Nilai Fisika dan Matematika *Cluster* 6 menduduki peringkat ke empat, sedangkan nilai Kimia menduduki peringkat ke enam (Tabel 5).

Cluster 8 yang menduduki peringkat ke enam memiliki 9.95% dari data (Tabel 2 dan Tabel 4). *Cluster* 8 menduduki peringkat ke 3 untuk nilai Kimia (Tabel 5). *Cluster* 8 memiliki kemampuan yang kurang di bidang Biologi, Fisika, dan Matematika dengan masing-masing peringkat ke 6, 7, dan 7 (Tabel 5). Secara keseluruhan, rata-rata nilai Biologi, Fisika, Kimia, dan Matematika *Cluster* 8 berada di bawah rata-rata (71.91 dari rata-rata 73.16), Tabel 4.

Cluster 7, 2, dan 1 merupakan 3 *cluster* dengan rata-rata nilai di bawah rataan keseluruhan. *Cluster* 7 menduduki peringkat ke tiga untuk nilai Biologi (Tabel 5), namun nilai yang lainnya di bawah rata-rata.

Daerah Asal

Secara keseluruhan dari semua *cluster*, bisa kita lihat bahwa pelamar dari Sumatera (1) paling banyak di *Cluster* 3, dan semakin *menurun* mengikuti turunnya rataan *cluster* (Tabel 6). Demikian juga dengan pelamar yang berasal dari Nusa Tenggara (5) dan Sulawesi (7) (Tabel 6).

Tabel 6 Persentase asal pelamar dalam setiap *cluster*

Cluster ke-	Asal Pelamar						
	1	*3*	*5*	*6*	*7*	*8*	*9*
3	45.18	46.19	3.05	0.00	4.57	1.02	0.00
9	28.70	69.57	1.74	0.00	0.00	0.00	0.00
5	21.61	76.88	0.50	1.01	0.00	0.00	0.00
4	17.92	79.72	0.94	0.47	0.47	0.47	0.00
6	18.11	78.19	2.06	0.82	0.82	0.00	0.00
8	14.29	81.48	2.12	1.59	0.00	0.53	0.00
7	10.66	86.03	0.37	1.47	0.74	0.74	0.00
2	13.73	84.51	0.00	1.06	0.35	0.00	0.35
1	11.70	87.23	0.53	0.53	0.00	0.00	0.00
n	19.17	77.67	1.16	0.84	0.79	0.32	0.05

Keterangan: **n** = data keseluruhan.

Sumatera memiliki persentase yang lebih besar dari persentase dia sendiri secara keseluruhan di *Cluster* 3, *Cluster* 9, dan *Cluster* 5 (Tabel 6). Hal ini menunjukkan bahwa, pelamar dari Sumatera banyak berada pada *cluster* yang memiliki rataan lebih tinggi.

Mayoritas anggota dari data berasal dari Jawa (3) sebesar 77.67% (Tabel 6), dengan persentase terkecil di *Cluster* 9. Secara keseluruhan persentase pelamar dari Jawa semakin meningkat mengikuti turunnya rataan *cluster* (Tabel 6). Terlihat bahwa pelamar yang berasal dari Jawa banyak berada di *cluster* yang memiliki rataan lebih rendah (*Cluster* 4, *Cluster* 6, *Cluster* 8, *Cluster* 7, *Cluster* 2, dan *Cluster* 1) (Tabel 6). Pelamar dari Luar Negeri 0.05% dari data berada hanya di *Cluster* 2 (Tabel 6).

Persentase pelamar yang diterima menunjukkan penurunan sebanding dengan penurunan nilai rataan *cluster*. Hal ini berlaku untuk daerah asal Sumatera, Jawa, dan Nusa Tenggara (Tabel 7). Untuk pelamar dengan daerah asal Kalimantan, Sulawesi, Irian Jaya, dan Luar Negeri hanya diterima untuk satu *cluster* tertentu (Tabel 7). Secara keseluruhan, pelamar yang terbanyak diterima adalah yang berasal dari Jawa yaitu sebesar 80.43%, bahkan seluruh pelamar *Cluster* 1 berasal dari Jawa (Tabel 7).

Tabel 7 Persentase pelamar yang diterima dalam setiap daerah asal

Cluster ke-	Asal Pelamar						
	1	*3*	*5*	*6*	*7*	*8*	*9*
3	54.12	14.60	66.67	0	100	100	0
9	17.65	11.68	16.67	0	0	0	0
5	10.59	15.09	0	0	0	0	0
4	3.53	16.30	0	0	0	0	0
6	5.88	16.30	0	0	0	0	0
8	5.88	9.49	16.67	0	0	0	0
7	2.35	9.49	0	100.00	0	0	0
2	0	4.87	0	0	0	0	100
1	0	2.19	0	0	0	0	0
n	16.63	80.43	1.17	0.20	1.17	0.20	0.20

Keterangan: **n** = data keseluruhan.

Putusan

Persentase putusan tidak diterimanya pelamar dari tiap *cluster* berbanding terbalik dengan rataan nilai keseluruhan (Tabel 8, dengan 0=tidak diterima, 1=diterima di IPB, A=diterima di Fakultas Pertanian, dan lain=diterima di fakultas selain Fakultas Pertanian). Kendati *Cluster* 6, *Cluster* 8, *Cluster* 7, *Cluster* 2, dan *Cluster* 1 memiliki rataan nilai keseluruhan di bawah rata-rata (Tabel 4), *cluster-cluster* tersebut masih memiliki persentase diterima (Tabel 8). Hal ini menunjukkan bahwa nilai pelamar bukan satu-satunya acuan dalam seleksi penerimaan mahasiswa baru.

Tabel 8 Detail putusan masing-masing *cluster* diurut berdasarkan rataan nilai.

Cluster ke-	0	1	
		A	lain
3	5.76	23.59	20.19
9	3.67	13.27	9.62
5	9.22	12.04	21.15
4	10.23	13.76	13.46
6	12.32	14.25	13.46
8	10.37	8.60	9.62
7	16.57	8.11	8.65
2	18.95	4.67	1.92
1	12.90	1.72	1.92
n	73.09	21.43	5.78

Keterangan: **n** = data keseluruhan.

Kategori Sekolah Asal Pelamar

Dari keseluruhan data terlihat persentase dari pelamar dalam setiap kategori SMA tersebar merata dalam tiap-tiap *cluster* (Tabel 9). Persentase pelamar dalam satu kategori SMA, semakin meningkat sebanding dengan rataan *cluster* (Tabel 10). Hal ini menunjukkan bahwa peluang untuk diterima dari masing-masing kategori bergantung kepada nilai.

Tabel 9 Persentase pelamar dari setiap *cluster* dalam satu kategori SMA

Cluster ke-	A+	A	A-	B+	B	B-	C+	C	C-	D
3	9.1	7.4	9.1	12.1	28.1	14.3	0	57.1	0.0	12.8
9	8.4	5.1	4.4	5.2	5.5	10.7	0	0.0	0.0	10.3
5	12.1	12.7	7.5	5.2	10.3	21.4	0	0.0	0.0	5.1
4	13.3	10.3	10.6	6.9	10.3	14.3	0	14.3	10.3	5.1
6	10.9	13.5	14.3	20.7	11.6	7.1	0	0.0	17.2	10.3
8	9.1	12.1	10.6	17.2	6.2	0.0	50	0.0	3.4	2.6
7	13.6	14.2	14.1	20.7	15.1	14.3	50	14.3	6.9	20.5
2	14.3	17.0	15.8	6.9	8.9	10.7	0	14.3	24.1	20.5
1	9.4	7.8	13.5	5.2	4.1	7.1	0	0.0	37.9	12.8
n	31.4	27.0	25.3	3.1	7.7	1.5	0.1	0.4	1.5	2.1

Keterangan: **n** = data keseluruhan.

Tabel 10 Persentase pelamar yang diterima dari setiap *cluster* dalam satu kategori SMA

Cluster ke-	A+	A	A-	B+	B	B-	C+	C	C-	D
3	18.83	15.44	27.40	38.46	51.22	16.67	0	100	0	41.67
9	13.45	11.03	12.33	15.38	12.20	16.67	0	0	0	8.33
5	15.25	14.71	13.70	0	9.76	25	0	0	0	0
4	14.35	15.44	17.81	0	7.32	8.33	0	0	0	0
6	13.00	17.65	12.33	23.08	9.76	16.67	0	0	0	8.33
8	8.97	11.76	5.48	23.08	4.88	0.00	0	0	0	0
7	9.87	8.09	6.85	0	2.44	0.00	0	0	0	25.00
2	4.04	4.41	2.74	0	2.44	16.67	0	0	0	8.33
1	2.24	1.47	1.37	0	0	0	0	0	0	8.33
n	43.64	26.61	14.29	2.54	8.02	2.35	0	0.20	0	2.35

Keterangan: **n** = data keseluruhan.

Dalam sebuah cluster, semakin baik kategori sebuah SMA, maka persentase pelamar yang diterima semakin tinggi (Tabel 11). Hal ini menunjukkan bahwa dari data terlihat kategori SMA berkontribusi terhadap diterima/tidaknya pelamar.

Tabel 11 Persentase pelamar yang diterima dari setiap kategori SMA dalam satu *cluster*

Cluster ke-	A+	A	A-	B+	B	B-	C+	C	C-	D
3	35.90	17.95	17.09	4.27	17.95	1.71	0	0.85	0	4.27
9	46.88	23.44	14.06	3.13	7.81	3.13	0	0	0	1.56
5	47.89	28.17	14.08	0	5.63	4.23	0	0	0	0
4	45.71	30.00	18.57	00	4.29	1.43	0	0	0	0.00
6	40.28	33.33	12.50	4.17	5.56	2.78	0	0	0	1.39
8	44.44	35.56	8.89	6.67	4.44	0.00	0	0	0	0
7	52.38	26.19	11.90	0	2.38	0.00	0	0	0	7.14
2	42.86	28.57	9.52	0	4.76	9.52	0	0	0	4.76
1	55.56	22.22	11.11	0	0	0	0	0	0	11.11
n	43.64	26.61	14.29	2.54	8.02	2.35	0	0.20	0	2.35

Keterangan: **n** = data keseluruhan.

5. KESIMPULAN DAN SARAN

Kesimpulan

Dari hasil percobaan ditemukan bahwa *clustering* terhadap data yang memiliki DBI minimal adalah ukuran *output* 9 dengan *learning rate* 0.9, penurunan *learning rate* 0.1, dan 5 iterasi yang menghasilkan DBI 53.472. Pelamar dengan pilihan pertama Fakultas Pertanian dari Sumatera banyak berada pada *cluster* yang memiliki rataan lebih tinggi (*Cluster* 3, *Cluster* 9, dan *Cluster* 5, dengan rataan nilai Biologi, Fisika, Kimia, dan Matematika dari masing-masing *cluster* (81.12, 77.50, dan 74.16). Berbeda dengan daerah asal Jawa, pelamar yang berasal dari Jawa banyak berada di *cluster* yang memiliki rataan lebih rendah (*Cluster* 4, *Cluster* 6, *Cluster* 8, *Cluster* 7, *Cluster* 2, dan *Cluster* 1, dengan rataan masing-masing 74.08, 73.09, 71.91, 70.04, 68.59, dan 67.93). Pelamar dari Luar Negeri tergolong pelamar dengan nilai rendah hanya berada di *Cluster* 2 dengan rataan 68.59.

Peluang untuk diterima dari masing-masing kategori bergantung kepada nilai, namun nilai pelamar bukan satu-satunya acuan dalam seleksi penerimaan mahasiswa baru. Kategori SMA juga berkontribusi terhadap diterima/tidaknya pelamar.

Saran

Penelitian selanjutnya dapat difokuskan untuk optimasi kombinasi nilai-nilai parameter algoritma SOM untuk memperoleh hasil yang optimal.

DAFTAR PUSTAKA

Demuth H, Beale M. 2003. *Neural Network Toolbox For Use with MATLAB®*. USA: The MathWorks, Inc.

Han J, Kamber M. 2001. *Data mining: Concepts and Techniques*. USA: Academic Press.

Jain AK, Dubes RC. 1988. *Algorithms for Clustering Data.* New Jersey: Prentice Hall Inc.

Kaski S. 1997. Data Exploration Using *Self organizing maps* [tesis]. Finlandia: Laboratory of Computer and Information Science, Department of Computer Science and Engineering, Helsinki University of Technology.

Larose DT. 2004. *Discovering Knowledge in Data: An Introduction to Data mining*. USA: John Wiley&Sons Inc.

Laurence F. 1994. *Fundamentals of Neural Networks.* New Jersey: Prentice Hall Inc.

Riyanti EF. 2005. Pengembangan Aplikasi Data Mining Menggunakan Metode Induksi Berorientasi Atribut (Studi Kasus: Data PPMB IPB) [skripsi]. Bogor: Departemen Ilmu Komputer, FMIPA-IPB.

Salazar GEJ, Veles AC, Parra MCM, Ortega LO. 2002. A Cluster Validity Index for Comparing Non-hierarchical Clustering Methods [terhubung berkala]. http://citeseer.ist.psu.edu/rd/salazar02cluster.pdf [10 Januari 2006]

Tan PN, Steinbach M, Kumar V. 2004. Introduction to Data Mining [terhubung berkala]. http://www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap2_data.ppt [23 Mei 2006].

Wisnujati I. 2006. Pembentukan Sistem Inferensi *Fuzzy* Mamdani dengan *Fuzzy C-Means* untuk Data Mahasiswa Baru IPB Tahun 2000-2004 [skripsi]. Bogor: Departemen Ilmu Komputer, FMIPA-IPB.