DATA MINING DENGAN TEKNIK *CLUSTERING* DALAM PENGKLASIFIKASIAN DATA MAHASISWA STUDI KASUS PREDIKSI LAMA STUDI MAHASISWA UNIVERSITAS BINA NUSANTARA

Lindawati
Jurusan Teknik Information, Fakultas Ilmu Komputer, Universitas Bina Nusantara
Jl. K.H. Syahdan No. 9, Kemanggisan – Palmerah, Jakarta Barat 11480
email: lindawati@binus.edu

Abstrak
Penggunaan teknik Data Mining *(DM)* clustering *berbeda dengan teknik-teknik DM yang lainnya, seperti* association rule mining *dan* classification *yang memerlukan tahapan training dan evaluasi. Teknik ini menggunakan metode* unsupervised learning *yang berarti DM tidak perlu melakukan training terlebih dahulu tapi bisa langsung menggunakannya untuk pengelompokan. Teknik ini masih jarang digunakan dibanding dengan teknik-teknik DM yang lain. Oleh karena itu, artikel ini berfokus pada pengembangan aplikasi DM dengan teknik* clustering *pada penelitian untuk mengklasifikasikan data prediksi lama studi masiswa di Universitas Bina Nusantara dengan menggunakan algoritma* Self Organizing Maps *dan pengujian keakuratan aplikasi DM dengan teknik* clustering*. Tahapan yang dilakukan dibagi menjadi tahapan analisa, perancangan, pengembangan dan pangujian aplikasi DM. Pada tahapan analisa dilakukan beberapa percobaan dalam mengklasifikasikan prediksi lama studi mahasiswa berdasarkan delapan atribut yang digunakan, yaitu: rata-rata Indeks Prestasi Kumulatif (rIpk), simpangan rata-rata Indeks Prestasi Kumulatif (srIpk), rata-rata jumlah SKS per Semester (rSksem), rata-rata jumlah SKS yang tidak lulus per semester (rSksemTL), jumlah SKS Kumulatif (skKum), jumlah SKS yang akan diambil pada semester keempat (sksYad), jumlah SKS yang wajib diambil (sksMin) dan hak SKS yang dapat diambil pada semester lima dst (hakSks). Hasil dari tahapan analisa ini digunakan sebagai acuan pada tahapan perancangan dan pengembangan aplikasi DM. Selanjutnya dilakukan pengujian terhadap aplikasi DM yang telah dibuat untuk mengetahui keakuratan pengklasifikasian dari aplikasi DM tersebut. Evaluasi dilakukan melalui beberapa variasi pengujian dengan menggunakan paramter-parameter jumlah data, jumlah iterasi, learning rate, radius, neighbourhood function dan urutan data. Dari pengujian-pengujian yang dilakukan dapat diketahui bahwa rata-rata kesalahan hasil klasifikasi prediksi lama studi yang diperoleh relatif kecil, kurang dari atau maksimal 5%.*
Kata Kunci: *Data Mining, clustering, Self-Organizing Maps, Prediksi Lama Studi Mahasiswa*

1. PENDAHULUAN

Banyaknya data yang dipunyai sebuah organisasi bisa menyebabkan kesulitan dalam pengklasifikasian data tersebut untuk kepentingan organisasi. Kegiatan pengklasifikasian yang dilakukan oleh manusia masih memiliki keterbatasan, terutama pada kemampuan manusia dalam menampung jumlah data yang ingin diklasifikasikan. Selain itu bisa juga terjadi kesalahan dalam pengklasifikasian yang dilakukan. Salah satu cara mengatasi masalah ini adalah dengan menggunakan *Data Mining* (DM) dengan teknik *clustering*.

Penggunaan teknik DM *clustering* berbeda dengan teknik-teknik *Data Mining* (DM) yang lainnya, seperti *association rule mining* dan *classification* yang memerlukan tahapan *training* dan evaluasi. Teknik ini menggunakan metode *unsupervised learning* yang berarti DM tidak perlu melakukan *training* terlebih dahulu tapi bisa langsung menggunakannya untuk pengelompokan. Teknik ini masih jarang digunakan dibanding dengan teknik-teknik DM yang lain.

Penelitian yang sudah dilakukan untuk mengklasifikasikan data prediksi lama studi masiswa di Universitas Bina Nusantara ini ditujukan untuk merancangan dan mengimplementasikan DM dengan teknik Clustering menggunakan algoritma *self organizing maps*. Tujuan lainnya adalah untuk menguji keakuratan *clustering* dengan menggunakan algoritma *self organizing maps*.

2. TINJAUAN PUSTAKA

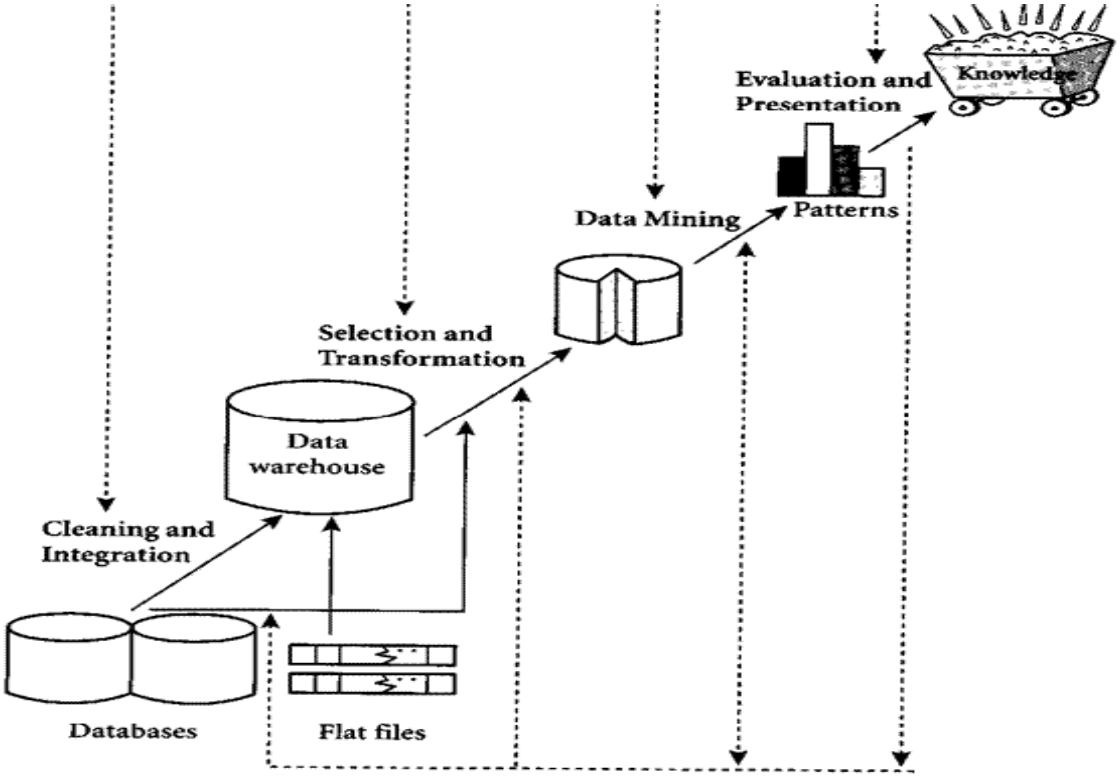
Data mining (DM) adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara *manual*. Patut diingat bahwa kata *mining* sendiri berarti usaha untuk mendapatkan sedikit barang berharga dari sejumlah besar material dasar. Karena itu DM sebenarnya memiliki akar yang panjang dari bidang ilmu seperti kecerdasan buatan (*artificial intelligent*), *machine learning*, statistik dan basis data. Beberapa teknik yang sering disebut-sebut dalam *literatur* DM antara lain : *clustering, classification, association rule mining, neural network,* dan *genetic algorithm*.

Perkembangan DM yang pesat tidak dapat lepas dari perkembangan teknologi informasi yang memungkinkan data dalam jumlah besar terakumulasi. Sebagai contoh, toko swalayan merekam setiap penjualan

barang dengan memakai alat POS (*point of sales*). Basis data penjualan tersebut. bisa mencapai beberapa GB (*GigaBytes*) setiap harinya untuk sebuah jaringan toko swalayan berskala nasional. Perkembangan internet juga punya andil cukup besar dalam akumulasi data. Tetapi pertumbuhan yang pesat dari akumulasi data itu telah menciptakan kondisi yang sering disebut sebagai "*rich of data but poor of information*" karena data yang terkumpul itu tidak dapat digunakan untuk aplikasi yang berguna. Tidak jarang kumpulan data itu dibiarkan begitu saja seakan-akan "kuburan data" (data tombs). DM mencoba untuk menyajikan informasi dari kumpulan data ini.

Karena DM adalah suatu rangkaian proses, DM dapat dibagi menjadi beberapa tahap. Tahap-tahap tersebut bersifat *interaktif* di mana pemakai terlibat langsung atau dengan perantaraan *knowledge base*. Tahap-tahap ini diilustrasikan di Gambar 1:

1. Pembersihan data (untuk membuang data yang tidak konsisten dan *noise*)
2. Integrasi data (penggabungan data dari beberapa sumber)
3. Transformasi data (data diubah menjadi bentuk yang sesuai untuk di-*mining*)
4. Aplikasi teknik DM
5. Evaluasi pola yang ditemukan (untuk menemukan yang menarik/bernilai)
6. Presentasi pengetahuan (dengan teknik visualisasi)



Gambar 1. Tahap-Tahap Data Mining

DM dengan teknik *clustering*, berbeda dengan teknik *association rule mining* dan *classification* dimana kelas data telah ditentukan sebelumnya. C*lustering* melakukan pengelompokan data tanpa berdasarkan kelas data tertentu. Bahkan *clustering* dapat dipakai untuk memberikan label pada kelas data yang belum diketahui itu. Karena itu *clustering* sering digolongkan sebagai metode *unsupervised learning*. Prinsip dari *clustering* adalah memaksimalkan kesamaan antar anggota satu kelas dan meminimumkan kesamaan antar kelas atau *cluster*. *Clustering* dapat dilakukan pada data yan memiliki beberapa atribut yang dipetakan sebagai ruang multidimensi.

Salah satu algoritma untuk membuat DM dengan teknik *clustering* adalah algoritma *Self-Orginizing Maps* (SOM) yang diperkenalkan oleh Prof. Teuvo Kohonen pada tahun 1982. SOM bekerja berdasarkan *competitive learning*, yaitu data-data (pada SOM disebut sebagai *neuron*) pada kumpulan data (dalam SOM disebut jaringan) saling berkompetisi satu sama lain untuk menjadi pemenang, dengan hasil berupa hanya satu data keluaran untuk setiap kelompok dalam setiap satuan waktu.

SOM diinspirasi oleh cara kerja otak manusia dalam menanggapi beberapa rangsangan sensorik yang diterima oleh panca indera. Neuron-neuron akan disesuaikan secara selektif dengan beberapa pola masukan dengan tujuan sebagai proses pelatihan kompetitif. Kemudian neuron-neuron akan diletakkan secara teratur dan terurut dalam suatu sistem koordinat yang berarti, yaitu dimulai dari neuron pemenang yang diikuti dengan

neuron-neuron lainnya yang memiliki kemiripan dengan neuron pemenang. Sehingga semakin mirip suatu neuron dengan neuron pemenang maka letaknya akan semakin dekat dengan neuron pemenang.

Hal itulah yang merupakan *topographic map* pada SOM. *Topographic map* adalah suatu pemetaan yang mempertahankan hubungan-hubungan di antara vektor-vektor masukan yang bersebelahan atau berdekatan. Inilah yang sering disebut sebagai *topology-preserving map* pada SOM.

3. METODE PENELITIAN

Tahapan yang dilakukan pada penelitian ini dibagi menjadi 4 tahapan, yaitu analisa, perancangan, pengembangan dan pengujian aplikasi DM. Hasil dari tahapan analisa akan mempengaruhi perancangan dan pengembangan aplikasi DM. Setelah pengembangan aplikasi selesai, dilanjutkan dengan tahapan pengujian aplikasi.

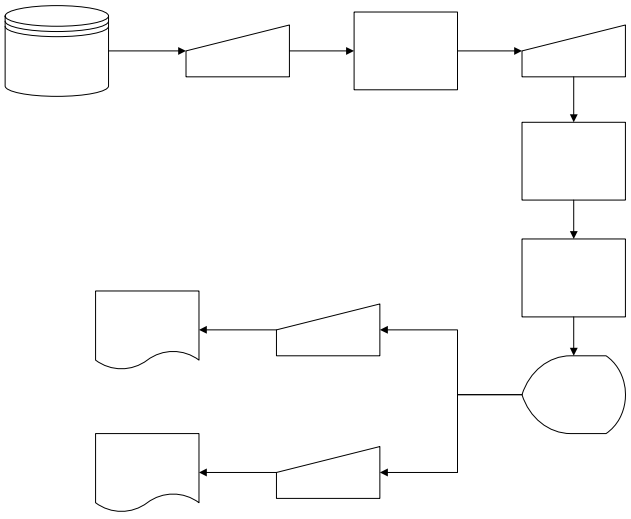
Tahapan analisa dilakukan untuk melakukan pemilihan *Decision Clasification* yang berguna untuk menandakan atau memberikan kriteria untuk kelompok-kelompok (cluster-cluster) yang terbentuk. Untuk mendapatkan suatu *decision classification* yang digunakan dalam pengklasifikasian prediksi lama studi, dilakukan percobaan dengan empat metode yang berbeda. Keempat metode tersebut dilakukan dengan menggunakan data 1024 mahasiswa yang telah diketahui lama studi yang sebenarnya dengan menggunakan lima atau delapan atribut. Dari masing-masing percobaan keempat metode tersebut dilakukan perbandingan antara kelompok prediksi lama studi yang diperoleh dengan kelompok lama studi yang sesungguhnya dengan menggunakan data 1024 mahasiswa secara acak. Dari rata-rata kesalahan maka metode yang digunakan sebagai decision classification adalah metode yang memiliki nilai atau rata-rata kesalahan lebih kecil dan juga menghasilkan map yang lebih baik, yaitu metode keempat.

Di tahapan ini juga ditentukan *atribute-atribute* yang digunakan untuk mengklasifikasikan prediksi lama studi mahasiswa Universitas Bina Nusantara. *Atribut-atribut* tersebut adalah sebagai berikut:

1. Rata-rata Indeks Prestasi Kumulatif (*rIpk*)
2. Simpangan rata-rata Indeks Prestasi Kumulatif (*srIpk*)
3. Rata-rata jumlah SKS per Semester (*rSksem*)
4. Rata-rata jumlah SKS yang tidak lulus per semester (*rSksemTL*)
5. Jumlah SKS Kumulatif (*skKum*)
6. Jumlah SKS yang akan diambil pada semester keempat (*sksYad*)
7. Jumlah SKS yang wajib diambil (*sksMin*)
8. Hak SKS yang dapat diambil pada semester lima dst (*hakSks*)

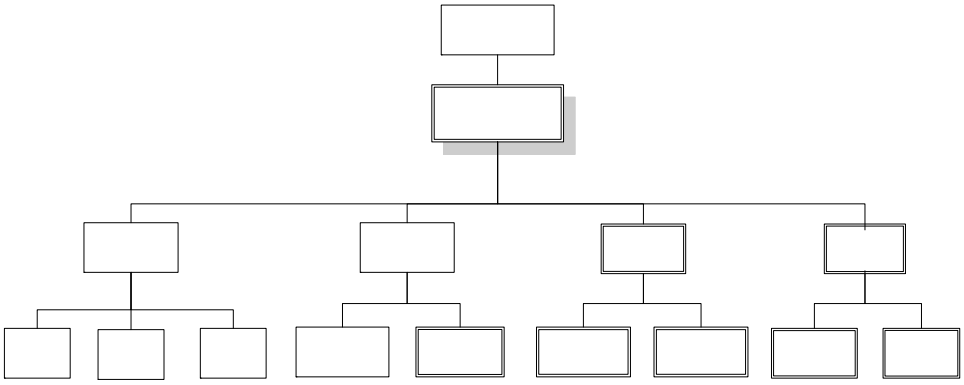
Berdasarkan atribut-atribut diatas, mahasiswa dikelompokan menjadi 5 kelompok yaitu kelompok pertama memiliki prediksi lama studi <= 4 tahun (warna biru), kelompok kedua memiliki prediksi lama studi > 4 tahun dan <= 5 tahun(warna hijau), kelompok ketiga memiliki prediksi lama studi > 5 tahun dan <= 6 tahun (warna jingga), kelompok keempat memiliki prediksi lama studi > 6 tahun dan <= 7 tahun (warna kuning) dan kelompok kelima memiliki prediksi lama studi > 7 tahun (warna merah). Kelompok-kelompok tersebut diwakilkan oleh warna-warna yang berbeda agar mudah membedakannya.

Dari hasil tahapan analisa tersebut, dimulailah tahapan perancangan dan pengembangan aplikasi. Aplikasi yang seterusnya disebut StepSOM dirancang dan dikembangkan dengan arsitektur sistem yang digambarkan pada gambar 2.



Gambar 2. Arsitektur Sistem

StepSOM bisa mengambil data dari basis data atau *data warehouse* dan dilengkapi dengan layar *login* sehingga hanya bisa diakses oleh pengguna yang berhak atau memiliki *login*. Setelah berhasil *login*, pengguna dapat menggunakan data baru yang diperoleh dari basis data atau *data warehouse* ataupun data yang telah ada sebelumnya (sudah diambil dari basis data atau *data warehouse* sebelumnya dan disimpan di file berekstensi dat) sebagai data masukan. Lalu pengguna melakukan inisialisasi *variable-variable* yang akan digunakan. Setelah itu dapat dilanjutkan dengan proses *training*. Proses *training* yang dilakukan adalah untuk memperoleh pengelompokan prediksi lama studi mahasiswa. Hasil dari proses *training* ini akan disimpan dalam suatu file berekstensi cod dan file berekstensi dbf.
Pengguna dapat melihat hasil dalam suatu *topographic map* (contohnya dapat dilihat digambar 1) yang ditampilkan dalam bentuk *rectangular grid* yang berdimensi *N* x *N* (besarnya dimensi *rectangular grid* ditentukan dari banyaknya data, apabila banyaknya data 100 maka *rectangular grid* yang dihasilkan 10x10). Pada *topographic map* tersebut akan dapat diketahui dengan mudah kelompok-kelompok mahasiswa berdasarkan prediksi lama studinya. Setiap kelompok memiliki satu warna sendiri yang akan membedakan kelompok tersebut dengan kelompok-kelompok lainnya. Hasil dalam bentuk *topographic map* disimpan ke dalam file berekstensi cod. Selain dalam bentuk *topographic map*, pengguna StepSOM juga dapat melihat tampilan laporan dalam bentuk diagram dan detail datanya. Detail data dari hasil proses pengelompokan disimpan di dalam file berekstensi dbf.
Menu-menu yang terdapat dalam StepSOM secara lengkap dapat dilihat pada gambar 3. Pada saat StepSOM diload untuk pertama kali dan juga pada saat StepSOM berada dalam keadaan *idle* (tidak ada satu dokumen pun yang sedang aktif pada layar MDI), sub menu *Training*, menu *Result* (termasuk sub menu *Map Visualization* dan *View As Grid*), dan menu *Report* (termasuk sub menu *Histogram* dan *Detail*) tidak dapat diklik atau *disabled*. Sub menu *Training* akan dapat diklik (*enabled*) jika sudah melakukan proses inisilisasi melalui sub menu *Initialization*. Sedangkan menu *Result* (termasuk sub menu *Map Visualization* dan *View As Grid*) dan menu *Report* (termasuk sub menu *Histogram* dan *Detail*) akan dapat diklik (*enabled*) jika sudah melakukan proses *training* melalui sub menu *Training*.



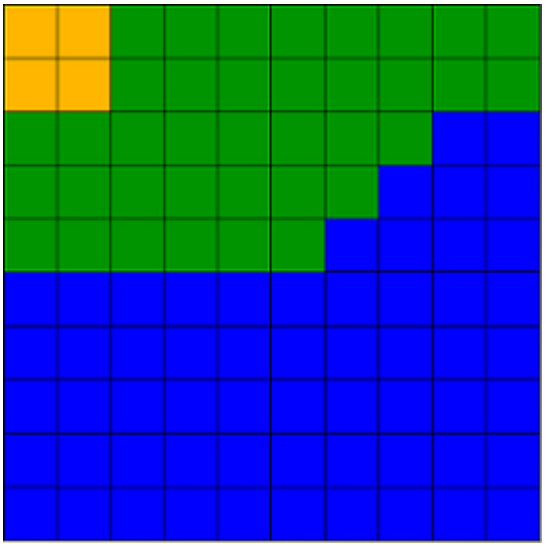
Gambar 3 Struktur Menu pada StepSOM

4. HASIL DAN PEMBAHASAN

Setelah aplikasi StepSOM selesai dikembangkan, dilakukan tahap pengujian. Tahapan ini dilakukan dengan beberapa kali percobaan terhadap aplikasi StepSOM dengan menggunakan 3 set data yang sudah tersedia dan tidak perlu mengambil dari basis data atau *data warebouse*. Tiga set data yang digunakan adalah 100 data acak, 1024 data acak dan 1024 data urut (terurut indeks prestasinya). Masing-masing set data dilakukan percobaan sebanyak 100 kali dengan melakukan variasi terhadap parameter-parameter berikut ini: banyaknya iterasi, dimensi *map*, *variable alpha*, *variable radius*, dan *neighbourhood function*.

4.1. Set Data 100 acak

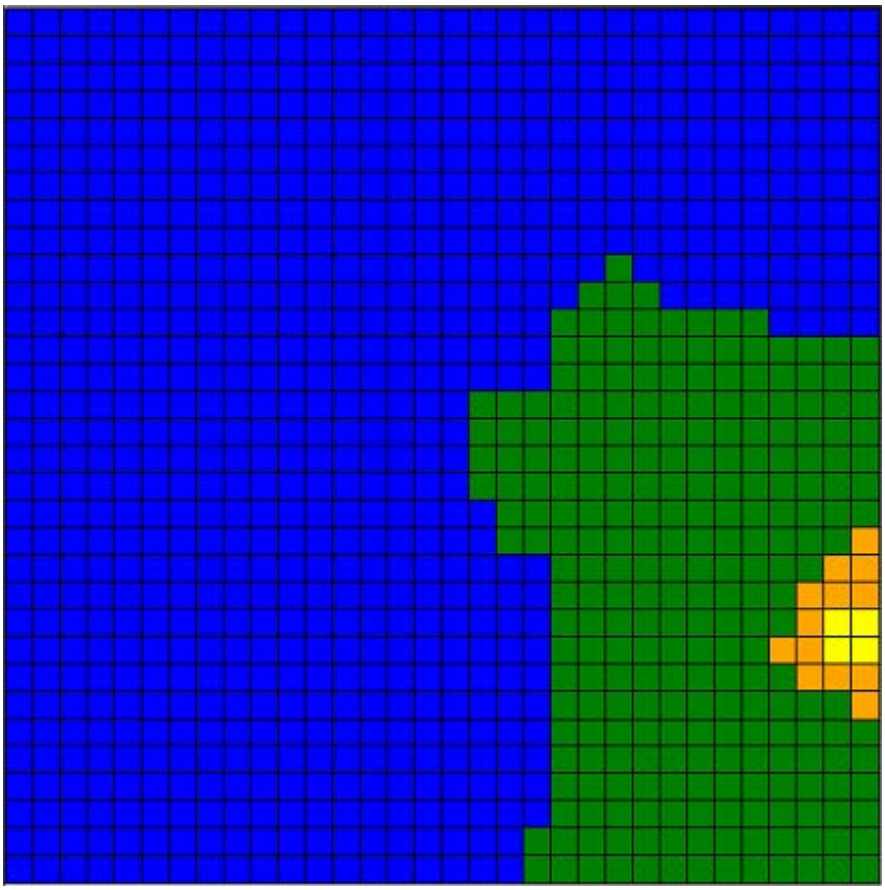
Dari set yang berisi 100 data acak, diperoleh rata-rata waktu penyelesaiannnya adalah 3.375 detik dan dihasilkan visualisasi *map* 10x10 seperti gambar 4. Dengan menggunakan set 100 data acak ini hanya dihasilkan 3 kelompok yaitu yang berwarna biru (kelompok ke-1), hijau (kelompok ke-2) dan kuning (kelompok ke-3), sedangkan kelompok dengan warna kuning (kelompok ke-4) dan merah (kelompok ke-5) tidak ada. Kelompok ke-1 mendominasi hasil pengelompokan (59%) dan kelompok ke-3 hanya mengisi sekitar 4 % dari keseluruhan set data, dan sisanya adalah kelompok ke-2.



Gambar 4. Map Visualization untuk 100 data acak

4.2. Set Data 1024 acak

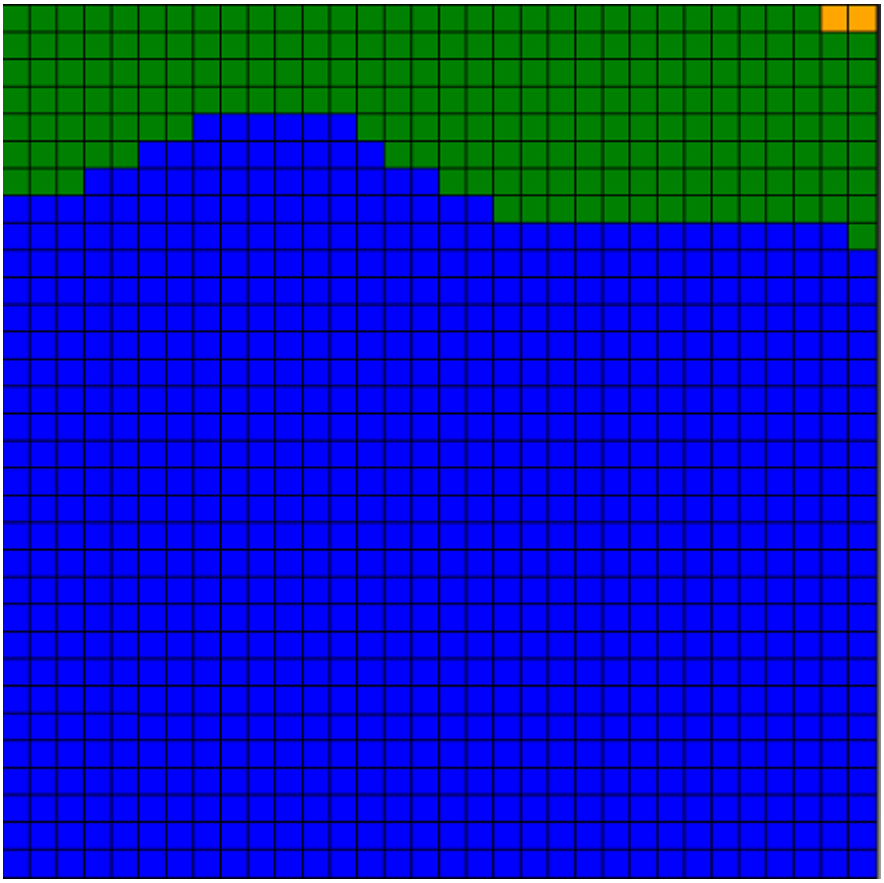
Dari set yang berisi 1024 data acak, diperoleh rata-rata waktu penyelesaiannnya adalah 232.845 detik dan dihasilkan visualisasi *map* seperti gambar 5. Dengan menggunakan set 1024 data acak ini dihasilkan 4 kelompok yaitu yang berwarna biru (kelompok ke-1), hijau (kelompok ke-2), kuning (kelompok ke-3) dan jingga (kelompok ke-4), sedangkan kelompok dengan merah (kelompok ke-5) tidak ada. Sama seperti pada set data 100 acak, hasil pengelompokan didominasi oleh kelompok ke-1, diikuti oleh kelompok ke-2, kelompok ke-3 dan kelompok ke-4.



Gambar 5. Map Visualization untuk 1024 data acak

4.3. Set Data 1024 Urut

Dari set yang berisi 1024 data urut, diperoleh rata-rata waktu penyelesaiannnya adalah 233.817 detik dan dihasilkan visualisasi *map* seperti gambar 6. Dengan menggunakan set 1024 data acak ini kembali dihasilkan 3 kelompok yaitu yang berwarna biru (kelompok ke-1), hijau (kelompok ke-2), dan kuning (kelompok ke-3), sedangkan jingga (kelompok ke-4) dan kelompok dengan merah (kelompok ke-5) tidak ada. Sama seperti pada set data 100 acak dan set data 1024 acak, hasil pengelompokan didominasi oleh kelompok ke-1, diikuti oleh kelompok ke-2, dan kelompok ke-3. Set data 1024 urut menggunakan data yang sama dengan set data 1024 acak, perbedaannya hanya pada set data 1024 urut, datanya diurutkan berdasarkan indeks prestasi sedangkan pada set 1024 acak tidak diurutkan. Walaupun datanya sama, apabila cara pengurutannya berbeda akan menghasilkan hasil yang berbeda.



Gambar 6. Map Visualization untuk 1024 data urut

4.4. Evaluasi Hasil Pengujian

Dari 3 set data tersebut dan berbagai variasi parameter-parameter yang sudah dilakukan pengujian, dapat diketahui:

1. Hasil yang diperoleh dengan menggunakan data yang nilai atribut-atributnya tidak diurutkan (acak) lebih baik daripada hasil yang diperoleh dengan menggunakan data yang nilai atribut-atributnya diurutkan.
2. Hasil yang diperoleh dengan menggunakan 1024 data lebih baik daripada hasil yang diperoleh dengan menggunakan 100 data.
3. Hasil yang diperoleh dengan menggunakan data acak dengan melakukan perubahan-perubahan parameter, diperoleh :
   i. *neighbourhood function* Bubble lebih baik daripada dengan menggunakan *neighbourhood function* Gaussian.
   ii. dengan menggunakan 100 data acak, hasil yang terbaik (memiliki nilai rata-rata jumlah kesalahan yang minimal) diperoleh dengan menggunakan *neighbourhood function* Bubble, *learning rate* 0.1, dan radius 4 ($\frac{d}{3}$; *d* adalah besarnya dimensi yang digunakan) atau dengan menggunakan *learning rate* 0.25 dan radius 4 ($\frac{d}{3}$; *d* adalah besarnya dimensi yang digunakan).

iii. dengan menggunakan 1024 data acak, hasil yang terbaik (memiliki nilai rata-rata jumlah kesalahan yang minimal) diperoleh dengan menggunakan *neighbourhood function* Bubble, *learning rate* 0.1, dan radius 16 ($\frac{d}{2}$; *d* adalah besarnya dimensi yang digunakan).

5. KESIMPULAN

Berdasarkan beberapa pengujian dengan 3 set data yang berbeda, dapat diperoleh kesimpulan sebagai berikut:

1. Self-Organizing Map memadai dan efektif untuk mengklasifikasikan prediksi lama studi mahasiswa dengan menggunakan parameter-parameter tertentu. Hal ini didasarkan pada besarnya rata-rata kesalahan dengan menggunakan neighbourhood function Bubble dan data yang atribut-atributnya tidak diurutkan (acak) yang termasuk kecil (<=5%).
2. Jumlah data dan jumlah iterasi yang dimasukkan mempengaruhi banyaknya waktu yang dibutuhkan untuk mengklasifikasikan prediksi lama studi.
3. Parameter-parameter learning rate, radius, neighbourhood function, dan juga urutan data akan mempengaruhi klasifikasi prediksi lama studi dan juga map yang dihasilkan.
4. Ditinjau dari nilai rata-rata kesalahan yang relatif kecil (<= 5%), aplikasi klasifikasi prediksi lama studi mahasiswa yang dibuat layak untuk digunakan.

6. DAFTAR PUSTAKA

Fausett, Laurene. (1994). Fundamentals of Neural Networks : Architectures, Algorithms, and Applications. Prentice Hall, New Jersey.

J. Gehrke, R. Ramakrishnan, V. Ganti (1998). "Rainforest: A framework for fast decision tree construction of large datasets". International Conf. Very Large Data Bases (VLDB).

J. Han, J. Pei, Y. Yin (2000). "Mining frequent patterns without candidate generation". ACM-SIGMOD International Conf. Management of Data (SIGMOD'00).

J. Han , M. Kamber (2001). Data Mining : Concepts and Techniques. Morgan Kaufmann.

Suparno, Haryono. Lindawati. Wijaya, Rusmin. Martana, Henry (2005). Penerapan Algoritma Self-Organizing Maps Dalam Pengklasifikasian Data Mahasiswa : Studi Kasus Prediksi Lama Studi Mahasiswa Universitas Bina Nusantara. LTB Jurusan Teknik Informatika Universitas Bina Nusantara, Edisi Desember 2005. Universitas Bina Nusantara, Jakarta.

U. Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy (1996). *Advances in Knowledge Discovery and Data Mining*, MIT Press.