PENGEMBANGAN ALGORITMA SEMUT DENGAN MENGGUNAKAN KONSEP *GLOBAL DESIRABILITY* DAN *GLOBAL FREQUENCY* UNTUK PENGELOMPOKAN DATA

Saiful Bukhori

Program Studi Teknik Elektro
Universitas Jember

ABSTRAK

Dalam penelitian ini dilakukan analisis penggunaan algoritma semut dalam domain data mining untuk pengelompokan data. Algoritma yang digunakan didasarkan modifikasi dan perbaikan terhadap beberapa algoritma semut yang pernah dikembangkan oleh para peneliti sebelumnya. Algoritma semut yang didesain dipengaruhi oleh empat parameter utama, yaitu ant desirability, ant frequency, pengatur informasi heuristic α dan pengatur nilai konsentrasi pheromone β. Hasil analisis penggunaan algoritma semut dalam domain data mining ini menunjukkan bahwa kompleksitas waktu terburuk dari algoritma adalah O(mn) dengan m dan n berturut-turut menyatakan jumlah atribut dan jumlah record dari data. Perangkat lunak yang telah berhasil didesain dan diimplementasikan dalam lingkungan sistem operasi Windows telah diuji coba dengan menggunakan berbagai konfigurasi nilai parameter yang mempengaruhi hasil pengelompokan. Hasil uji coba menunjukkan prosentase jumlah aturan yang dikelompokkan dengan benar berkisar antara 93,48% - 97,00%, dikelompokkan salah berkisar antara 3,00% - 6,52% dan tidak dapat dikelompokkan 0%.

Kata kunci: data mining, pengelompokan data, algoritma semut, ant desirability, ant frequency

ABSTRACT

This research tries to analyse the application of ant algorithm in data mining for data clustering. The algorithm used was the modification and improvement of the algorithm that was developed by previous researchers. The ant algorithm designed was influenced by four main parameters, ant desirability, ant frequency, heuristic information (α) and pheromone concentration (β). The analysis showed that the worst time complexity of this algorithm was O(mn), where m and n represent number of attributes and number of data recorded, respectively. The software that was designed and implemented on the Windows operating system was tested using a number of parameter configurations. Result of the test showed that true clustering in the range of 93.48% - 97.00%, false clustering in the range of 3,00% - 6,52% range, and 0% unclustering.

Keywords: data mining, data clustering, ant algorithm, ant desirability, ant frequency

PENDAHULUAN

Data mining adalah kegiatan eksplorasi dan analisis secara otomatis terhadap suatu basis data yang besar dengan tujuan menemukan pola-pola dan aturan-aturan yang tersembunyi dan mempunyai makna bagi penggunanya. Dalam proses eksplorasi basis data tersebut, *data mining* membutuhkan algoritma yang berfungsi sebagai alat penggalinya.

Pengertian *data mining* tersebut mendasari diperlukannya alat penggali dalam hal ini adalah algoritma yang mempunyai kemampuan untuk menganalisis basis data yang besar sehingga dapat menghasilkan suatu informasi yang dapat menunjang proses pembuatan keputusan yang dapat dipertanggungjawabkan. Algoritma semut yang didasarkan pada pengamatan koloni semut yang saat ini terus dikembangkan merupakan pilihan dalam penelitian ini dengan harapan dapat membantu dalam pengembangan *data mining*.

Penelitian ini ditekankan pada analisis penggunaan algoritma semut pada *data mining* untuk pengelompokan data dengan menggunakan konsep *global desirability* dan *global frequency*.

Berdasarkan hasil penelitian ini dapat dirancang suatu perangkat lunak yang mampu memberikan informasi tentang pengelompokan data dari sebuah basis data yang berukuran besar, sehingga perangkat lunak ini dapat digunakan sebagai alat penunjang proses pengambilan keputusan.

Data Mining

Data mining yang juga dikenal dengan *knowledge discovery in database* (KDD) merupakan metode pencarian terhadap informasi yang bernilai dan tersembunyi dalam suatu basis data yang sulit atau bahkan tidak mungkin untuk ditemukan dengan menggunakan mekanisme *query* standard atau teknik statistik klasik (Anda, 1999).

Data mining dalam dunia bisnis disebut sebagai bagian dari *business intelligence* (BI) yang merupakan alat bantu bagi perusahaan untuk berkompetisi menguasai pasar sehingga mampu secara cepat memantau kecenderungan pasar. DSS (*decision support system*), *data warehouse*, OLAP (*on-line analytical processing*) dan *data mining* merupakan metode-metode penting dalam *business intelligence* ini (Joseph, 1996).

Data mining menggunakan teknik yang berbeda untuk menemukan pola data dan ekstraksi informasi. Pada umumnya yang sering digunakan adalah *query tools*, teknik statistik, visualisasi, *on-line analytical processing* (OLAP), *case-based learning* (*k-nearest neighbor*), *decision trees*, *association rules*, *neural networks*, atau *genetic algorithms*.

Jenis *data mining* didasarkan manfaatnya dapat digolongkan ke dalam klasifikasi, estimasi, prediksi, *affinity grouping*, pengelompokan (clustering), dan deskripsi (Joseph, 1996).

Konsep Dasar Algoritma Semut

Perilaku serangga sosial diarahkan untuk mempertahankan kehidupan keseluruhan koloninya, bukan individu-individu tunggal dari koloni tersebut. Perilaku ini menarik perhatian banyak ilmuwan karena tingginya tingkat struktur koloni yang dapat dicapai, terutama apabila dibandingkan dengan kesederhanaan individu-individu anggota koloni tersebut. Perilaku yang penting dan menarik dari semut adalah kemampuannya untuk menemukan jalur terpendek antara sumber makanan dan sarang mereka (Marco, 2001). Selama semut berjalan dari tempat sumber pangan ke sarang dan sebaliknya aktifitas semut antara lain adalah sebagai berikut:

a. Semut meninggalkan di atas jalannya suatu zat yang disebut *pheromone*, dengan demikian membentuk jejak *pheromone*
b. Semut dapat mencium bau *pheromone*, ketika memilih jalan
c. Jejak *pheromone* membuat semut mampu menemukan jalan kembali ke sumber pangan atau sarangnya
d. Pergerakan semut tidak terlalu banyak, bergantung pada keadaannya dan apakah dia melihat makanan di arah depannya atau tidak. Semut hanya akan melangkah

ke depan, belok kiri, belok kanan, atau diam

e. Memori semut berisi kondisinya atau disebut *chromosome*; *chromosome* ini terdiri dari gen-gen yang membawa informasi berupa apa yang ia kerjakan dan apa yang akan dikerjakan selanjutnya

Secara eksperimental telah terbukti bahwa perilaku mengikuti jejak *pheromone* ini, apabila diterapkan oleh sebuah koloni semut, mengarah kepada ditemukannya jalur terpendek.

METODOLOGI

Desain algoritma semut untuk *data mining* pengelompokan data dalam penelitian ini didasarkan konsep *global desirability* dan *global frequency*. Konsep *global desirability* dan *global frequency* akan membangkitkan semua solusi yang dalam kasus ini berupa kelompok-kelompok data sesuai dengan pola data berdasarkan ciri-ciri atribut yang penting dari suatu data atau berdasarkan analisis dari atribut-atribut tersebut. Agar hal ini dapat dicapai, dibutuhkan 2 jenis *agent*, yaitu *agent global desirability* dan *agent global frequency*.

Struktur algoritma semut untuk *data mining* dalam pengelompokan data dalam penelitian ini dirancang terdiri dari 4 proses utama yaitu proses pemasukan data, proses pengkodean, proses pembentukan jejaring *pheromone*, dan proses pengelompokan data baru.

Proses Pemasukan Data

Proses pemasukan data terdiri dari pembacaan data set yang akan dikelompokkan dan masukan dari pemakai. Proses pembacaan data set akan menghasilkan informasi tentang *field data*, *record data*, *value* (nilai) dari *record data* tersebut pada *field* tertentu, jumlah dan nama kelompok yang dikelompokkan dan dimensi data. Masukan data dari pemakai berupa parameter-parameter yang dibutuhkan oleh algoritma untuk proses pengelompokan data tersebut. *Pseudocode* proses pemasukan data dapat dirancang seperti dalam teks algoritma berikut:

```
Jml_Ant_des = Jml Ant global
desirability;
Jml_Ant_Freq = Jml Ant global
Frequency;
Alpha = Parameter informasi heuristic;
Betha = Parameter nilai pheromone;
FOR (0 ≤ i ≤ Data_Set → RecordCount)
BEGIN
IF (nama_field == "Class") THEN
Namaclus=Data_Set→Field
nama("Class");
int k = 0;
Bool next=true;
WHILE((next==true)and(k<jumlahclus))
BEGIN
IF (anamaclus[k] == namaclus)THEN
Next= false; k++;
END
FOR (0<j< DataSet→FieldCount - 1)
BEGIN
Temp = 0;
IF (nama_field !="Class) THEN
nonkelas =
Dataset→Field→field[i][j];
Temp++;
END
END
```

Proses Pengkodean

Masukan proses pengkodean berasal dari *array nonclus* yang merupakan *array* yang berisi data yang belum dikelompokkan, yang dihasilkan pada proses pemasukan data. Elemen-elemen *array nonclus* yang merupakan nilai dari atribut-atribut untuk masing-masing *record* menentukan terbentuknya simpul dan busur pada jejaring *pheromone*. Pembentukan simpul dan busur ditentukan oleh agen *global desirability*, sedangkan konsentrasi *pheromone* yang menentukan kualitas aturan yang telah dibentuk ditentukan agen *global frequency*.

Proses pembentukan aturan dilakukan dengan analisis atribut-atribut berdasarkan ada atau tidak adanya hubungan antara atribut dengan nilai yang dimodelkan sebagai vektor biner. Atribut ke-i yang dilambangkan dengan A_i menyatakan simpul ke i. Nilai ke-j dari atribut ke-i yang disimbolkan dengan V_{ij} menyatakan simpul ke-j dari atribut ke-i. Jika ada busur dari simpul i ke simpul j, maka elemen (A_i, V_{ij}) ditandai dengan "1", dan sebaliknya, bila tidak ada busur dari simpul i ke simpul j, maka elemen (A_i, V_{ij}) ditandai dengan "0". *Pseudocode* proses pengkodean

seperti dalam teks algoritma berikut:

```
FOR (j=0; j <= FieldCount - 1; j++)
BEGIN
FOR (i=0; i <= RecordCount; i++)
BEGIN  //Pembentukan term_ij
value =Table1->Fields[j]->AsString;
term[i][j] = value;
IF(term_ij == term_i-1 j) THEN code = 1;
ELSE IF (term_ij != term_i-1 j) THEN code = 0;
END
END
```

Proses Pembentukan Jejaring *Pheromone*

Masukan proses pembentukan jejaring *pheromone* adalah pola data biner. Dari pola data tersebut, dapat dibentuk model himpunan solusi yang mungkin yang merupakan *subset* dari himpunan solusi yang dimodelkan dengan ruang berdimensi n. Pembentukan ruang berdimensi n dilakukan dengan diawali dari ruang dengan dimensi 2 kemudian dimensi 3, dimensi 4 sampai dengan dimensi n dengan cara menarik sudut-sudut dimensi sebelumnya. Kode biner '0' ditambahkan terhadap prefik dari kode bagian sudut luar dan kode '1' ditambahkan terhadap prefik dari kode sudut bagian dalam.

Simpul himpunan solusi yang dimodelkan dengan ruang berdimensi n tersebut menunjukkan adanya aturan yang telah dibentuk oleh agen semut yang berfungsi membentuk segmen berdasarkan *global desirability*, sedangkan busur antara simpul-simpul tersebut menunjukkan lintasan yang diikuti oleh agen semut yang berfungsi membentuk segmen berdasarkan *global frequency*.

Proses pembentukan kelompok dimulai dengan pembentukan struktur dalam bentuk aturan *IF (condition) THEN (clus)*. Masing-masing semut mulai dengan aturan tanpa *term* dalam *antecedent* dan menambah satu *term* pada setiap waktu untuk aturan yang sedang dibentuk. Aturan yang sedang dibentuk, dibangun oleh semut sesuai dengan bagian yang diikuti oleh semut tersebut. Demikian juga, pemilihan *term* untuk ditambahkan ke bagian aturan yang sedang dibentuk sesuai dengan pemilihan bagian yang berlangsung dan akan diperluas ke semua bagian yang memungkinkan. Pemilihan *term* untuk ditambahkan bergantung pada fungsi *heuristik* dan jumlah kumpulan *pheromone* di masing-masing *term*. Berdasarkan konsep tersebut di atas, *pseudocode* untuk proses pembentukan jejaring *pheromone* seperti dalam teks algoritma berikut:

```
FOR(j=0; j<= RecordCount; j++)
BEGIN
FOR (i=0; i<= FieldCount - 1; i++)
BEGIN  // Rule Construction
value = Table1->Fields[j]; //
Menentukan Nilai Pheromone
IF (ARRAY [i][j] == value ) THEN
THO = THO + 1; //Menentukan Nilai
LOG2 (Kelas)
LENK= (( k ** 0.001)-1)/0.001;
LOGKLS[k] = LENK/LENB; //
Menentukan Nilai INFO τij
  namakelas = Table1-
>FieldByName("CLASS");
  IF (ARRAY [i][j] == value  &&
RECORD[i] == namakelas)THEN
    FREKTHO = FREKTHO + 1;
 FREKW = FREKTHO / THO;
    LENA = (( FREKW ** 0.001)-
1)/0.001;
    LENB = 0.693;
    LOGINFO = LENA/LENB;
    INFOTHO = FREKW * LOGINFO;
    INFOTHOK = INFOTHOK + INFOTHO;
    HEUR = INFOTHO/INFOTHOK;
//Menentukan Nilai Heuristik
    PROB1=(HEUR**Alpha)*(THO *
Beta);//Tentukan Prob. term
    PROB[i] = PROB1;
    IF (PROB[i][J] >= PROB[i-1][j])
THEN
      PROB = PROB[i][j];
      PROBATURAN [i] = PROBATURAN +
PROB;
      PROBATURAN = PROBATURAN[i];
      Aturan = Term[i][j];//Rule
berdasar penambahan termij
END
END
FOR(j=0; j<= RecordCount; j++)
BEGIN
IF (PROBATURAN[j] >= PROBATURAN[j-
1])THEN
PROBATURANKLAS = PROBATURAN[j];
END
```

Proses Pengelompokan

Proses pengelompokan digunakan untuk menentukan kelompok yang sesuai dengan aturan yang terbaik yang ditemukan oleh algoritma, sedangkan jumlah kelompok yang telah ditemukan dalam proses pemasukan data digunakan sebagai acuan

untuk menentukan jumlah aturan yang diambil dari proses pembentukan jejaring pheromone yang dilakukan sebelumnya. Proses peng-*update*-an kelompok dilakukan dengan cara menarik semua aturan yang berada dalam dimensi yang terluar menjadi dimensi di bawahnya satu demi satu sampai mencapai dimensi aturan terbaik. Dari konsep tersebut dapat dibuat *pseudo code* seperti dalam teks algoritma berikut:

```
FOR (k = 0; k< JmlKls; k++)
BEGIN
FOR (j = 0; j<= RecordCount; j++)
BEGIN
IF (PRATURAN[j] >= PRATURAN[j-1])
PRATURANKLAS = PRATURAN[j]
PRRULECLUS[k]=((jmlclus-
k)/jmlclus)*PRATURANKLS
END
END
KELAS[k] = PRAILITASRULECLASS[k]
```

Uji Coba Perangkat Lunak

Uji coba perangkat lunak dilakukan dalam lingkungan sistem operasi Windows yang dijalankan dengan komputer PC Pentium III 667 MHz dengan RAM berkapasitas 128 MB dan *harddisk* sebesar 10 GB. Dalam penelitian ini, terdapat empat jenis data yang digunakan dalam uji coba yaitu *Dermatologi*, *German Credit*, *Car* dan *Nursery* dengan spesifikasi sebagaimana ditunjukkan pada Tabel 1. Data ini dapat diperoleh di http://www1.ics.uci.edu/~mlearn/MLSummary.html.

Tabel 1 Kasus yang Diujicobakan

Data Set	Jumlah Record	Jumlah Atribut Kategorikal	Jumlah Kelas
Dermatologi	366	33	6
German Credit	1000	13	2
Car	1210	6	4
Nursery	12960	8	5

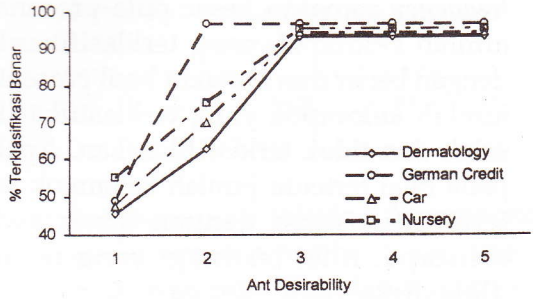
HASIL DAN PEMBAHASAN

Hasil uji coba direpresentasikan pada Gambar 1 hingga Gambar 4. Dari Gambar 1 sampai dengan Gambar 4 dapat diidentifikasi beberapa catatan penting berikut:

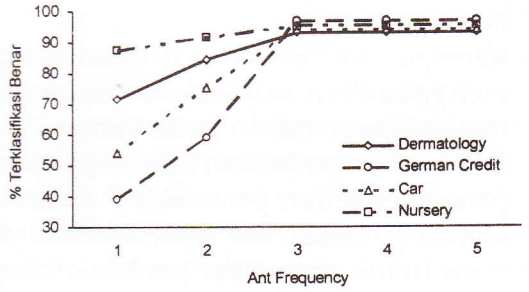
1. Keempat set data yang diujikan dengan menggunakan skenario 1 dan 2 yang hasilnya ditunjukkan pada Gambar 1 dan Gambar 2 menunjukkan bahwa semakin besar nilai parameter *ant desirability* dan *ant frequency* semakin besar pula prosentase jumlah kelompok yang terklasifikasikan dengan benar dan semakin kecil prosentase jumlah kelompok yang terklasifikasikan salah dan tidak terklasifikasikan. Apabila pada nilai tertentu jumlah kelompok yang terklasifikasikan dengan benar sudah mencapai nilai tertinggi yang mampu diklasifikasikan dengan benar oleh perangkat lunak, maka penambahan nilai parameter tidak akan menambah nilai prosentase yang terklasifikasikan dengan benar.
2. Keempat set data yang diujikan dengan menggunakan skenario 3 dan 4 yang hasilnya ditampilkan pada Gambar 3 dan Gambar 4 menunjukkan bahwa apabila nilai parameter α dan paramater β sama atau seimbang, maka akan diperoleh nilai yang besar untuk prosentase jumlah kelompok yang terklasifikasikan dengan benar, dan nilai yang kecil untuk prosentase jumlah kelompok yang salah terklasifikasikan dan tidak terklasifikasikan.

Waktu yang dibutuhkan untuk eksekusi program dapat dibagi menjadi 3 bagian yaitu waktu pembacaan data, waktu proses pengelompokan, dan waktu untuk pengelompokan data baru. Tabel 2 menunjukkan waktu eksekusi program. Tabel 2 menunjukkan bahwa proses pengelompokan membutuhkan waktu yang lebih lama bila dibandingkan dengan proses yang lain. Hal ini disebabkan karena pada proses tersebut adalah proses pembelajaran yang membentuk aturan dari masing-masing kelompok, sehingga pada proses ini bergantung pada jumlah *record*, jumlah atribut dan jumlah aturan kelompok yang dikenali oleh perangkat lunak. Tahapan pembacaan set data hanya dipengaruhi oleh jumlah *record* dan jumlah atribut yang dimiliki oleh dataset yang diujikan, sedangkan tahapan pengelompokan data baru hanya membutuhkan waktu untuk

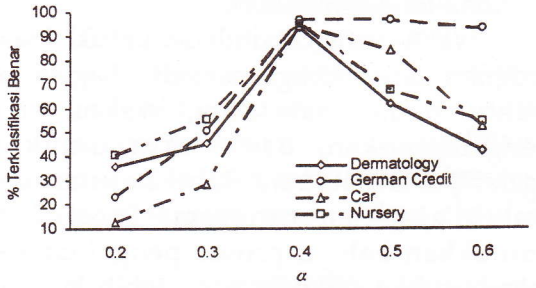
menetapkan solusi kelompok yang terbaik dari data yang dimasukkan sehingga hanya tergantung dari jumlah aturan yang dikenali sebelumnya.



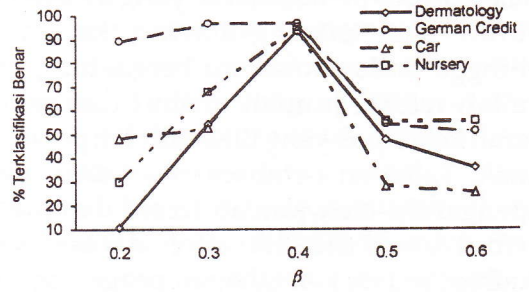
Gambar 1 Pengaruh *Ant Desirability* pada *Ant Frequency* = 500 , α = 0,40 dan β = 0,40



Gambar 2 Pengaruh *Ant Frequency* (x 100) pada *Ant Desirability* = 2 , α = 0,40 dan β = 0,40



Gambar 3 Pengaruh Parameter α pada β =0,4 *Ant Desirability* = 2 , dan *Ant Frequency*= 500



Gambar 4 Pengaruh Parameter β pada α = 0,4 *Ant Desirability* = 2 , dan *Ant Frequency*= 500

Tabel 2 Kinerja Waktu Eksekusi Pengelompokan Data

Data Uji Coba	Waktu (detik)		
	Pembacaan Set Data	Proses Pengelom-pokan	Pengelom-pokan Data Baru
Dermatology	24	52	1
German Credit	25	56	1
Car	22	51	1
Nursery	120	170	1

KESIMPULAN

Beberapa kesimpulan dapat diperoleh dari hasil penelitian yang dilakukan sehubungan dengan analisis penggunaan algoritma semut dalam *data mining* untuk pengelompokan data:

- Berdasarkan setting nilai parameter data uji coba yang digunakan dalam penelitian ini, hasil uji coba menunjukkan bahwa semakin besar nilai parameter *ant desirability* dan *ant frequency* semakin besar pula persentase jumlah kelompok yang dikelompokkan dengan benar, semakin kecil besar persentase jumlah aturan yang salah dikelompokkan maupun yang tidak dikeompokkan, akan tetapi apabila sudah dicapai nilai terbesar dari persentase jumlah kelompok yang dikelompokkan dengan benar yang mampu dilakukan oleh perangkat lunak, maka penambahan nilai kedua parameter tersebut tidak akan mampu lagi menambah jumlah persentase yang dikelompokkan dengan benar.
- Dalam uji coba data set *German Credit* diperoleh persentase nilai yang dikelompokkan dengan benar terbesar yaitu 97,00%, sedangkan pada data set *Dermatology* persentase nilai yang dikelompokkan benar yaitu 93,47%, pada data set *Car* prosentase nilai dikelompokkan dengan benar yaitu 94,28% dan pada data set *Nursery* persentase nilai dikelompokkan benar yaitu 95,38%.
- Berdasarkan persentase jumlah aturan yang dapat dikelompokkan secara benar, sesuai dengan kategori kelompok yang ada dalam data uji coba, hasil uji coba

menunjukkan bahwa hasil yang diperoleh dengan menggunakan nilai parameter yang memberikan hasil terbaik untuk setiap skenario uji coba memberikan persentase keberhasilan pengelompokan (97,00% untuk data uji coba *German credit*, 93,47% untuk data uji coba *Dermatology*, 94,28 % untuk data uji coba *Car* dan 95,38% untuk data uji coba *Nursery*).

DAFTAR PUSTAKA

Joseph, B.P. 1996, *Datamining with Neural Networks: Solving business problem form Application Development to Decision Support.* McGraw-Hill. New York

Anda, C. 1999. *Datamining Techniques in Supporting Decision Making*. Master Thesis. Universiteit Leiden. http://www.Ainet-sp.si.vti.bin.shtml.dll/ education .htm

Cover, T.M. & Thomas, J.A. 1991. *Element of Information Theory*. John Wiley & Sons. New York

Marco, D. Andrea, R. & Christian, B. 2001. HC-ACO: *The Hyper-Cube Framework for Ant Colony Optimization*. The 4th Metaheuristic International Conference (MIC'2001). Porto. Portugal