Reduksi Data Menggunakan Algoritma Genetika

Ali Ridho Barakbah, Entin Martiana
Politeknik Elektronika Negeri Surabaya – ITS
Kampus ITS Sukolilo Surabaya 60111
Email: ridho@eepis-its.edu; entin@eepis-its.edu

ABSTRAK

Data yang besar pada data mining membutuhkan beban komputasi yang tinggi. Salah satu cara untuk mengurangi hal tersebut diperlukan suatu reduksi data sehingga jumlah data semakin kecil. Permasalahan yang timbul pada reduksi data biasanya adalah apakah jumlah data setelah mengalami proses reduksi dapat mewakili jumlah data awal yang besar. Paper ini mengajukan suatu pendekatan baru dalam mereduksi data dengan menggunakan algoritma genetika. Jumlah data hasil reduksi direpresentasikan menjadi kromosom. Untuk memilih agar data hasil reduksi mendekati data sesungguhnya, digunakan fungsi fitness berupa jumlah jarak terbesar dalam kromosom. Kombinasi dari masing-masing kromosom harus dioptimasi terlebih dahulu dengan pendekatan pemodelan optimasi pada Travelling Salesman Problem. Untuk mengukur tingkat presisi dari reduksi data ini, dilakukan dengan mengakumulasi kedekatan masing-masing centroid hasil cluster dari data reduksi terhadap centroid hasil cluster dari data sesungguhnya. Hasil eksperimen menunjukkan tingkat presisi dari pendekatan reduksi data yang diajukan pada paper ini dengan menggunakan Algoritma Genetika. Kinerja dari pendekatan reduksi data pada paper ini akan dibandingkan dengan reduksi data dengan menggunakan algoritma Density-Based Multiscale Data Condensation.

Kata Kunci: *Reduksi data, algoritma genetika, data mining*

ABSTRACT

A large data set in data mining needs a high computational cost. One of approaches to reduce the cost is to reduce the data so that the number of data becomes smaller. The problem that usually emerges regarding data reduction is what can the data after reduction be representative for whole data. This paper proposes a new approach for data reduction using genetic algorithm. A total number of reduced data is represented as chromosome. To choose the data which can represent the whole data is used a fitness function which is the highest accumulated distance in the chromosomes. The combination of each chromosomes must be optimized at first with making an approach for optimization model in travel salesman problem. To measure the precision of condensation is to accumulate the closeness between each centroids from clustering result of reduced data to centroids from clustering result of the whole data. The experiment result performs the precision of a new approach of data reduction using genetic algorithm proposed in this paper. The performance of the proposed approach ini this paper will be compared to an algorithm of data condensation using Density-Based Multiscale Data Condensation.

Keywords: *data reduction, genetic algorithm, data mining.*

1. PENDAHULUAN

Permasalahan yang populer dari Data Mining dan Data Warehousing adalah biaya untuk penyimpanan data, dimana data yang harus disimpan bisa mencapai ukuran terabyte. Proses penggalian data dengan ukuran gigabyte yang kecil ketika diproses dengan sebuah metode machine learning seringkali sudah membutuhkan perangkat keras dan algoritma secara paralel. Penyelesaian untuk permasalahan ini adalah dengan pemilihan sub himpunan yang kecil dari data untuk proses learning. Di satu sisi yang lain data seringkali berisi data yang redundan Dengan demikian akan lebih baik data yang besar diwakili oleh sub himpunan kecil dengan pola dari data asli direpresentasikan dalam data reduksi.

Penerapan sederhana dari reduksi data adalah teknik random, yaitu pemilihan data reduksi dilakukan secara random. Terdapat beberapa pemilihan data reduksi secara statistik, dimana setiap anggota data mempunyai peluang untuk dipilih sebagai sampel, yaitu random sampling, stratified sampling, dan peepholing. Metode reduksi yang sederhana tidak dapat diterapkan pada data nyata dengan noise. Penerapan algoritma random menghilangkan informasi dari data-data yang tidak terpilih dalam proses reduksinya. Algoritma reduksi harus menyertakan informasi dari semua data dalam proses reduksinya.

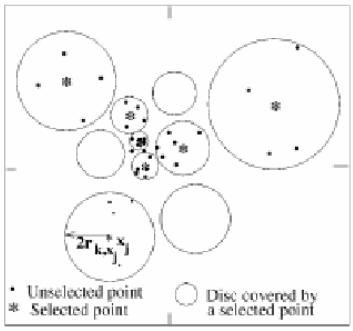
Beberapa skema penelitian untuk reduksi data dibangun dari penerapan klasifikasi secara umum dan aturan *k*-NN secara khusus. Keefektifan dari data reduksi terukur dalam keakuratan hasil klasifikasi. Metode yang pertama kali adalah penyingkatan data dengan aturan *k*-NN adalah CNN diinspirasikan oleh Hart. Beberapa algoritma lain yang termasuk dalam reduksi NN dan algoritma penyingkatan secara iteratif. Matrik pembobotan

 ISBN 979-545-027-1 @2006 ITS

kesamaan secara asimetrik lokal (LASM) diterapkan pada reduksi data dan menunjukkan kinerja yang bagus dibanding metode berbanding *k*-NN. Selain itu terdapat algoritma reduksi data dan model bervariasi berbasiskan neural network didiskusikan dalam. Bagaimanapun juga metode reduksi yang berbasiskan pada klasifikasi sangat spesifik terhadap model dan masalah klasifikasi yang digunakan. Reduksi data yang ditampilkan dengan metode vektor kuantisasi klasik menggunakan sebuah himpunan yang berisi vektor kode yang meminimumkan error kuantisasi.

Kelompok lain dari metode reduksi data adalah metode yang berbasiskan pada ukuran kepadatan data dengan mempertimbangkan fungsi kepadatan dari data lebih bertujuan untuk penyingkatan daripada meminimumkan error kuantisasi. Metode ini tidak menyertakan learning dalam prosesnya dan deterministik (dengan input yang sama akan memberikan output yang tetap) [6].

Metode Density Based Multiscale Data Consensation (DBMDC) [5] dikembangkan dari kelompok ini. Algoritma ini bekerja mengurangi data dengan melakukan pada skala yang berbeda. Dengan skala yang berbeda diharapkan algoritma efisien dalam memperkirakan error kepadatan dan sesuai dengan distribusi data. Selain penggunaan skala yang berbeda, dalam algoritma ini digunakan beberapa skala dengan harapan informasi yang ada akan diwakili dalam data reduksi.



Gambar 1. Ilustrasi reduksi data dengan DBMDC

Algoritma reduksi data meliputi perkiraan kepadatan pada suatu titik, mengurutkan titik-titik berdasarkan kriteria kepadatan, memilih suatu titik berdasarkan daftar urutan, dan pemangkasan semua titik yang berada dalam lingkaran dengan radius tertentu yang berbading terbalik dengan kepadatan pada titik tersebut. Sebuah metode non parametrik yang berfungsi memperkirakan probabilitas kepadatan dalam algoritma ini digunakan *k*-NN. Algoritma ini bekerja dengan langkah-langkah sebagai berikut:

Himpunan $B_N = \{x_1, x_2, ..., x_N\}$ sebagai data set inputan. Pilih nilai integer positif k .

1. Untuk tiap titik $x_i \in B_N$, hitung jarak k^{th} nearest neighbor dari x_i pada B_N . Tandai dengan r_{k,x_i} .
2. Pilih titik $x_j \in B_N$ yang mempunyai nilai r_{k,x_j} terkecil dan letakkan dalam himpunan E .
3. Hapus semua titik dari B_N yang berada dalam lingkaran radius $2r_{k,x_j}$ yang berpusat di x_j dan titik-titik yang tersisa di set sebagai B_N .

2. ALGORITMA GENETIKA UNTUK REDUKSI DATA

Algoritma genetika dikembangkan oleh John Holland [3]. Algoritma genetika adalah algoritma yang memodelkan proses seleksi alamiah biologi dari rekombinasi genetika, mutasi dan seleksi natural untuk membangkitkan solusi suatu permasalahan. Algoritma genetika digunakan untuk permasalahan maksimasi untuk mendapatkan solusi global optimal.

Gambaran umum dari algoritma genetika dasar adalah sebagai berikut [4]:

1. *t*=0
2. Inisialisasi populasi *P(t)*
3. Hitung fitness dari *P(t)*
4. *t*=*t*+1
5. Jika kondisi terpenuhi, langsung ke langkah
6. Dapatkan *P(t)* dari proses seleksi pada *P(t-1)*
7. Lakukan kawin silang (cross-over) pada *P(t)*
8. Lakukan mutasi pada *P(t)*
9. kembali ke langkah 3
10. Solusi terbaik didapatkan

2.1 Representasi kromosom

Representasi kromosom pada paper ini adalah data reduksi. Jika *n* adalah jumlah data yang telah direduksi, maka jumlah kromosom pada individu adalah sebanyak *n* kromosom.

Sedangkan jumlah gen dalam setiap individu adalah sebanyak *n* gen dimana setiap kromosom hanya mempunyai 1 gen. Representasi kromosom pada paper ini bersifat kombinatorial parsial, dengan kemungkinan nilai yaitu 1..*N* dimana *N*=jumlah data sebelum direduksi.

2.2 Pembangkitan Populasi

Pembangkitan populasi dilakukan secara acak dengan batasan kombinatorial parsial pada kromosom. Nilai hasil pembangkitan populasi berarti data yang tidak terkena proses reduksi.

2.3 Fungsi Fitness

Secara umum suatu cluster yang baik adalah yang mempunyai kohesi internal dan isolasi eksternal [2]. Suatu cluster yang ideal adalah cluster yang mempunyai varians dalam cluster yang kecil dan varians antar cluster yang besar [1].

Keterpisahan yang besar antar cluster akan membuat jarak data pada suatu cluster ke data pada cluster yang lain akan jauh.

Berangkat dari sini kami membuat suatu formulasi untuk mencari fungi fitness. Jika data keseluruhan itu direduksi sedemikian rupa sehingga jumlah datanya menjadi lebih sedikit, maka data hasil reduksi yang baik akan relatif identik dengan jumlah total jarak antar data-data tersebut yang semakin jauh. Semakin besar total jarak antar data-data reduksi berarti data-data reduksi semakin baik. Ini dikarenakan dengan tersebarnya data-data hasil reduksi, maka keseimbangan data-data untuk mewakili data keseluruhan semakin baik. Sehingga fungsi fitness yang dipakai pada paper ini adalah jumlah jarak pada data-data reduksi dalam kromosom.

Konsekuensi dari dipakainya fungsi fitness yang berupa jumlah jarak pada data-data reduksi yang direpresentasikan kromosom, maka urutan penempatan data-data reduksi harus diset tidak berfungsi. Persoalan ini sama dengan permasalahan Travelling Salesman Problem (TSP) dengan fungsi minimasi.

Katakanlah n adalah jumlah kromosom. Kalau n berjumlah sedikit, maka persoalan ini dapat diselesaikan dengan permutasi n. Dalam setiap kombinasi, rangkaian kromosom dihitung nilai fitness-nya, sehingga didapatkan $n!$ kombinasi nilai fitness. Selanjutnya adalah diambil nilai yang paling minimal.

Kalau n berjumlah banyak, maka penyelesaian dengan permutasi akan mempertinggi waktu komputasi. Dalam paper ini, kami menggunakan Algoritma Genetika untuk menyelesaikan disfungsi dari urutan penempatan data-data reduksi.

2.4 Seleksi

Seleksi dilakukan dalam rangka untuk mendapatkan calon induk yang baik. Karena induk yang baik akan menghasilkan keturunan yang baik. Sehingga semakin tinggi nilai fitnes suatu individu maka semakin besar kemungkinannya untuk terpilih. Dalam tahap seleksi ini dilakukan dengan menggunakan mesin *roullete*.

2.5 Kawin Silang

Pada paper ini kami tidak melakukan kawin silang. Ini dikarenakan kasus yang kami angkat adalah bersifat kombinatorial parsial sehingga perubahan allele pada gen dari pasangannya tidak diperlukan.

2.6 Mutasi

Mutasi dilakukan untuk mengekploitasi ruang solusi. Pada paper ini kami melakukan mutasi dengan cara replacement (penggantian). Jika N adalah jumlah data keseluruhan, n adalah jumlah data hasil reduksi, $G=\{g_1, g_2, ..., g_n\}$ adalah allele-allele pada gen dan p adalah allele pengganti untuk memutasi allele gen sebelumnya, maka p adalah hasil pembangkitan random $1..N$ dan $p \neq G$.

2.7 Elistisme

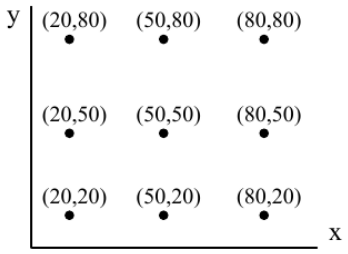
Elitism digunakan untuk mengikutkan individu-individu induk yang terbaik ke populasi baru dengan cara menggantikan individu-individu turunan yang terjelek [7], sehingga mencegah kehilangan solusi terbaik. Pada paper ini kami memakai sistem ranking. Kalau jumlah individu dalam populasi adalah sebanyak p, maka terdapat p individu induk dan p individu turunan. Dengan sistem ranking, populasi induk dan populasi turunan digabung sehingga berjumlah $2p$ individu. Selanjutnya setiap individu tersebut dihitung nilai fitness-nya dan diranking yang terbaik. Kemudian diambil sebanyak p yang terbaik untuk menjadi populasi baru.

3. SKENARIO EKSPERIMEN

3.1 Dataset

Dalam paper ini kami mengaplikasikan kasus reduksi data pada data yang well-separated (terpisah baik). Ini kami lakukan untuk dapat menganalisa lebih mudah apakah data-data reduksi yang dihasilkan nanti bisa representatif terhadap data keseluruhan.

Untuk membuat data yang well-separated, paper ini membangkitkan dataset dengan distribusi data normal random. Untuk tujuan percobaan, kami menggunakan 2 dimensi dataset. Lalu kami menentukan 9 posisi titik acuan, sebagaimana yang ada pada Gambar 2, untuk pembangkitan dataset secara random.



Gambar 2. Posisi titik-titik acuan untuk pembangkitan dataset berdistribusi normal random

Jumlah titik-titik acuan tersebut dibangkitkan secara random dengan minimal 6 titik acuan dan maksimal 9 titik acuan. Mengenai posisinya juga diambil secara random. Selanjutnya, pada masing-masing titik acuan, kami bangkitkan n data secara random pada radius r. Dengan pembangkitan data seperti ini, kami dapat membuat ribuan kombinasi dataset yang well-separated.

 ISBN 979-545-027-1 @2006 ITS

3.2 Parameter eksperimen

Pada paper ini, kami melakukan rangkaian percobaan untuk r=5 dan r=10. Untuk masing-masing r, kami melakukan percobaan dengan n={10, 15, 20, 25, 30}. Untuk masing-masing r dan n, kami melakukan 100 kali percobaan dan menghitung rata-rata kinerja dengan Proximity Index. Setiap percobaan, kami mencoba melakukan reduksi data dengan Algoritma Genetika yang kami ajukan pada paper ini, dan juga kami bandingkan dengan algoritma DBMDC.

3.3 Proximity Index

Proximity Index dipakai untuk mengukur sejauh mana kedekatan centroids dari data-data reduksi terhadap centroids dari data keseluruhan. Untuk penghitungan Proximity Index (PI) didapatkan sebagai berikut. Jika k adalah jumlah cluster, C={c_1, c_2,…, c_k} adalah centroids dari data-data hasil reduksi, dan D={d_1, d_2,…, d_k}, adalah centroids dari data keseluruhan, maka:

$$PI = \min \sum_{i=1}^{k} |C - D| \qquad (1)$$

Proximity Index akan mengungkapkan bahwa data-data hasil reduksi yang baik akan mempunyai PI yang kecil dikarenakan centroids dari data-data reduksi tersebut dekat dengan centroids dari data keseluruhan.

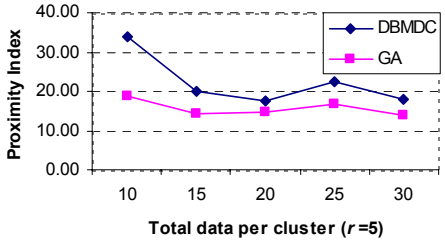
4. HASIL EKSPERIMEN

Setelah melakukan serangkaian percobaan, didapatkan tingkat kinerja dari reduksi data menggunakan Algoritma Genetika (GA) dengan algoritma DBMDC, seperti pada Tabel 1.

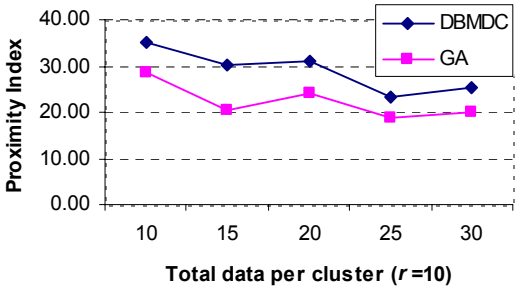
Tabel 1. Tabel perbandingan PI reduksi data dengan algoritma DBMDC dan GA

r	*n*	DBMDC	GA
5	10	33.89	18.92
5	15	20.15	14.44
5	20	17.73	14.66
5	25	22.25	16.67
5	30	17.84	13.78
10	10	35.18	28.70
10	15	30.26	20.48
10	20	31.22	24.28
10	25	23.39	18.77
10	30	25.51	19.83

Pada Tabel 1 terlihat bahwa PI yang dihasilkan oleh GA lebih kecil dibandingkan dengan DBMDC. Gambar 3 dan 4 memperlihatkan grafik perbandingan PI untuk masing-masing algoritma dengan r=5 dan r=10.



Gambar 3. Perbandingan PI untuk r=5.



Gambar 4. Perbandingan PI untuk r=10.

Pada Gambar 3 dan 4 terlihat bahwa PI yang dihasilkan oleh GA lebih kecil dibandingkan dengan DBMDC untuk semua r dan n. Dengan demikian, secara eksperimental penyelesaian reduksi data dengan memakai Algoritma Genetika yang kami usulkan dalam paper ini lebih baik dibandingkan dengan algoritma DBMDC untuk kasus dataset yang well-separated.

5. References

1. Barakbah, A.R., and Arai, K. Reversed pattern of moving variance for accelerating automatic clustering. *EEPIS journal*, p.15-21, Vol. 9, Number 2, EEPIS-ITS, Surabaya, December 2004.
2. Cowgill, M.C., and Harvey, R.J. A Genetic Algorithm Approach to Cluster Analysis. *Computers & Mathematics with Applications*, vol. 37, pp. 99-108, 1999.
3. Goldberg, D. Genetic Algorithms in Search, Optimization and Machine Learning. *Addison-Wesley*, Massachusetts, USA, 1989.
4. Maulik, U., and Bandyopadhyay, S. Genetic Algorithm Based Clustering Technique. *Pattern Recognition*, vol. 33, no. 9, pp. 1455-1465, 2000.
5. Mitra, P; Murthy, C.A. and Sankar K.P. Density-Based Multiscale Data Condensation. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol 24, No 6, 2002.
6. Palmer, C. R., and Faloutsos, C. Density Biased Sampling: An Improved Method for Data Mining and Clustering. *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD-2000)*, 2000.
7. Ying X., Tay L.P. Genetic Algorithm Based K-Means Fast Learning Artificial Neural Network. *17th Australian Computer Society Conference on Artificial Intelligence*, Lecture Notes in Artificial Intelligence (LNAI), 2004.