ANALISIS DAN IMPLEMENTASI KLASTERISASI MENGGUNAKAN FAST GENETIC K-MEANS ALGORITHM (FGKA)

Ni Putu Aryanti Kamadeni[1], Moch Arif Bijaksana[2], Sri Widowati[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Klasterisasi (clustering) merupakan salah satu fungsionalitas data mining yang digunakan untuk melakukan pengelompokan data ke dalam suatu kelas atau cluster. Prinsip dasar klasterisasi adalah mengelompokkan objek pada suatu kelas yang memiliki kemiripan sangat besar dengan objek lain pada kelas yang sama (similarity), tetapi sangat tidak mirip dengan objek pada kelas lain (dissimilarity). Terdapat beberapa teknik klasterisasi antara lain: metode Partisi (K-means Clustering), metode Hierarki (Divisive and Agglomerative Clustering), metode Density-Based (DBSCAN), dan sebagainya.

Pada tugas akhir ini, digunakan metode Partisi dengan algoritma FGKA (Fast Genetic K-means Algorithm) yang merupakan penggabungan antara algoritma Kmeans dan algoritma Genetika. Metode pengklasteran dengan menggunakan Kmeans sangat banyak digunakan untuk mengelompokkan data dengan similaritas yang tinggi. Akan tetapi K-means memiliki kelemahan dalam penentuan titik pusat inisial klaster yang dilakukan secara random sehingga sering kali menyebabkan terjebak pada lokal optimal dan hasil pengklasteran menjadi tidak optimal. Untuk lebih mengoptimalkan penentuan titik pusat dan dengan waktu yang seminimal mungkin maka digunakanlah algoritma FGKA (Fast Genetic Kmeans Algorithms). FGKA merupakan algoritma yang dikembangkan dari algoritma GKA (Genetic K-means Algorithm) yang diusulkan oleh Yi Lu pada tahun 2004. Algoritma ini selalu menghasilkan konvergensi pada global optimal. FGKA dan GKA mampu menghindari lokal optimal akan tetapi FGKA berjalan lebih cepat dibandingkan GKA. Dalam tugas akhir ini juga akan dilakukan perbandingan evaluasi hasil klasterisasi yang dihasilkan oleh perangkat lunak klasterisasi menggunakan metode K-means.

Kata Kunci : clustering, GKA, K-means, FGKA, Fast Genetic K-means Algorithm

Abstract

Clustering is one of data mining functionalities which is used to group data into classes or clusters. The basic principle of clustering is to group the object into cluster which has many similarities with other object in the same cluster and many dissimilarities with other object in different cluster. There are several clustering techniques, such as Partition method (K-means Algorithm), Hierarchical method (Divisive and Agglomerative Clustering), Density-Based method (DBSCAN), etc.

This final project is implemented Partition method with FGKA (Fast Genetic K-means Algorithm) which is merged by K-means algorithm and Genetic algorithm. K-Means algorithm is often used to group data that have many similarities. Nevertheless, K-Means has a weakness on determining centroid intial cluster point which is done randomly so that it causes K-Means trapped on local optimal and the result of clustering cannot be optimal.

To make the best of centroid point determining and minimal time, algoritma FGKA can be used. FGKA is algorithm which is developed from GKA algorithm proposed by Yi Lu in 2004. This algorithm always converges to a global optimum eventually. FGKA and GKA can avoid local optimum, but FGKA runs faster than GKA. This final project also compares the result of clustering evaluation from a clustering software with K-Means method.

Keywords : clustering, GKA, K-Means, FGKA, Fast Genetic K-Means Algorithm

1. Pendahuluan

1.1 Latar belakang

Data Mining merupakan salah satu bidang yang berkembang pesat karena besarnya kebutuhan akan nilai tambah dari database skala besar sebagai tuntutan dari pertumbuhan teknologi informasi. Dimana data mining itu sendiri adalah serangkaian proses untuk menggali nilai tambah berupa pengetahuan yang selama ini tidak diketahui secara manual dari kumpulan data [8]. *Clustering* merupakan salah satu fungsionalitas dari data mining yang digunakan dalam mengekstrak pengetahuan yang berguna untuk mendapatkan pola yang menarik dari volume data yang banyak dan dimensi data yang besar. Dimana *clustering* mengklasifikasikan data kedalam kelas atau *cluster* yang memiliki persamaan dan perbedaan yang dibawa oleh masing-masing atribut data. Klaster yang baik adalah klaster yang memiliki persamaan (*similarity*) intraklaster yang tinggi dan perbedaan (*dissimilarity)* antarklaster yang tinggi.

Saat ini banyak aplikasi yang menggunakan klasterisasi dalam pemecahan masalahnya sehingga banyak pendekatan yang ditawarkan dengan algoritmanya masing-masing , diantaranya adalah metode Partisi (K-means), Hierarki, Fuzzy C-means, dll. *Clustering* dengan metode K-means yang dikembangkan oleh Mac Queen pada tahun 1967, sangat terkenal dengan kemampuannya untuk mengklaster data yang besar dan dapat menangani data outlier. K-means merupakan metode pengklasteran yang memisahkan data kedalam *k* kelompok yang berbeda artinya sebelum dilakukan klasterisasi maka user harus menentukan jumlah *k* pastisi yang diinginkan. Selain itu pendekatan umum dari klasterisasi adalah menemukan titik pusat klaster yang merepresentasikan tiap klaster. Oleh karena itu K-means juga melakukan penentuan titik pusat klaster yang dibangkitkan dengan cara random. Hanya saja dalam penentuan titik pusat tersebut K-means masih sangat sensitif. K-means akan mampu menemukan titik pusat yang tepat apabila pembangkitan awal titik pusat yang dilakukan dengan random tersebut mendekati solusi akhir pusat klaster begitu juga sebaliknya. Jika awal titik pusat jauh dari solusi akhir pusat klaster maka kemungkinan besar hasil klasterisasinya menjadi tidak tepat. Dari keterangan tersebut diketahui bahwa K-means hanya dapat mencapai local optimal saja, belum mampu mencapai global optimalnya.

Selain algoritma K-means, beberapa peneliti juga menentukan klaster dengan algoritma genetika (Genetic Algorithm, disingkat GA). Dimana ide dasarnya adalah untuk mensimulasikan proses evolusi dari seleksi alami dan mengembangkan solusi dari satu generasi ke generasi berikutnya. Jika dibandingkan dengan K-means maka akan menjadi sangat kontras karena GA tidak sensitif pada inisialisasi awal dan selalu konvergen pada wilayah global. Hanya saja biaya komputasi menggunakan GA ini mahal untuk aplikasi yang luas. Ketidakefisienan waktu pada GA ini disebabkan karena GA menggunakan crossover operator yang membutuhkan waktu yang lama untuk menghasilkan kromosom anak yang valid dari kromosom induknya. Crossover operation sangat rumit dan membutuhkan perulangan yang tidak sedikit untuk menghasilkan kromosom legal. Selain itu juga membutuhkan biaya mahal pada perhitungan

fungsi fitnessnya. Oleh karena itu Khrisna dan Murty mencoba menggabungkan kekuatan alami GA dan kesederhanaan dari K-means menjadi algoritma *Genetic K-means Algorithm* (GKA) [9]. Dimana operator crossover pada GA digantikan dengan operator K-Means yang mengambil langkah serupa dengan *K-Means Algorithm* sebagai operator pencarian. Dan dari hasil penelitian mereka diperoleh hasil bahwa algoritma GKA lebih cepat dibandingkan menggunakan algoritma genetika murni karena GKA selalu menghasilkan konfergensi pada wilayah global.

Begitu juga dengan FGKA (*Fast Genetic K-means Algorithm)* yang merupakan pengembangan dari algoritma GKA dimana keduanya mampu menghasilkan konfergensi pada wilayah global [11]. Hanya saja performansi algoritma FGKA jauh lebih cepat dibandingkan dengan GKA. Oleh karena itu, hal tersebutlah yang menjadi alasan pemilihan algoritma FGKA dalam memecahkan kasus *clustering* sehingga diharapkan dapat diperoleh hasil klaster yang optimal dan dengan performansi semaksimal mungkin.

1.2 Perumusan masalah

Dengan mengacu pada latar belakang masalah di atas, maka permasalahan yang akan dibahas dan diteliti adalah :

1. Bagaimana menerapkan algoritma *FGKA (Fast Genetic K-means Algorithm)* pada metode *Clustering*.
2. Bagaimana performansi Algoritma *FGKA (Fast Genetic K-means Algorithm)* pada *Clustering* sebagai suatu metode dalam menentukan *cluster* dengan efektif dan menghasilkan *cluster* yang lebih optimal dibandingkan dengan menggunakan metode yang sejenis dan standar lainnya.

1.3 Tujuan

Berdasarkan pada masalah yang telah diidentifikasikan di atas, maka tujuan Tugas Akhir ini adalah:

1. Mengimplementasikan metode *Clustering* dengan *FGKA (Fast Genetic K-means Algorithm).*
2. Menganalisis hasil klasterisasi yang dihasilkan oleh perangkat lunak *Clustering* menggunakan *FGKA (Fast Genetic K-means Algorithm)* dan membandingkannya dengan hasil klasterisasi menggunakan metode K-means, yang mana parameter pembandingnya adalah akurasi (rata-rata error) dan *time performance* .

1.4 Metodologi penyelesaian masalah

Metode yang digunakan dalam penyelesaian tugas akhir ini adalah :

1. Studi Literatur
   Mencari referensi dan sumber-sumber lain yang layak yang berhubungan dengan *data mining*, *Clustering*, *Genetic Algorithm* dan *FGKA (Fast Genetic K-means Algorithm).*
2. Pendalaman Materi

Mempelajari konsep clustering dan algoritma *Fast Genetic K-means Algorithm* sehingga dapat menentukan tujuan yang ingin dicapai berdasarkan parameter -parameter inputan.

3. Perancangan dan Implementasi
   Merancang program dengan perancangan terstruktur dan mengimplementasikan hasil perancangan menggunakan bahasa pemrograman Delphi 7.
4. Analisis dan Evaluasi
   Melakukan pengujian perangkat lunak dengan menganalisa performansi metode Klasterisasi dengan *Fast Genetic K-means Algorithm* berdasarkan parameter input berupa jumlah klaster yang diinginkan (*k*), ukuran populasi (M), jumlah maksimum generasi (G), dan besarnya probabilitas mutasi yang digunakan dalam penentuan klaster dari masing– masing data.
5. Penyusunan Laporan Tugas Akhir
   Menyusun laporan hasil analisa yang dirangkum ke dalam sebuah buku Laporan Tugas Akhir.

5. Penutup

5.1 Kesimpulan

Dari uji kinerja dan analisis yang telah dilakukan pada bab IV terhadap 4 jenis data normal yang telah teruji, 3 jenis data 2 dimensi dan 4 jenis data ekstrim yang membandingkan metode clustering menggunakan FGKA dan K-Means maka dapat diambil kesimpulan sebagai berikut :

1. FGKA memiliki kemampuan lebih baik dibandingkan dengan K-means karena FGKA cenderung menghasilkan error rate , TWCV (Total Within Cluster Varian) yang lebih kecil dan TBCV (Total Between Cluster Varian) yang lebih besar, baik untuk data yang belum dinormalisasi maupun yang sudah dinormalisasi.
2. Hasil klaster dan akurasi klaster pada FGKA lebih stabil dibandingkan dengan K-Means.
3. Proses klasterisasi pada FGKA relatif lebih lama dibandingkan K-Means dan dari segi kompleksitas algoritmanya, FGKA sangat kompleks dibandingkan dengan K-Means.
4. FGKA maupun K_Means kurang mampu menangani pengklasteran data *Different Size* (Berbeda ukuran) , *Different Density* (Berbeda Kepadatan) dan data *Non Globular* (Tidak Globular) karena keduanya memiliki karakteristik dasar yang sama yaitu hanya terbatas pada perkiraan centroid dari data yang tersebar.
5. FGKA dapat digunakan untuk mengklaster semua jenis data extrim baik yang seluruh datanya sama maupun pada data yang sama pada tiap dimensinya.
6. FGKA Clustering akan menghasilkan hasil klaster yang lebih baik jika digunakan pada data-data yang globular.

5.2 Saran

1. Untuk memperoleh hasil klaster yang lebih baik, lakukan percobaan sebanyak mungkin dan cari rata-rata hasil yang paling optimal dengan melihat perbandingan masing-masing parameter inputan.
2. Untuk pengembangan metode Clustering, gunakan parameter input yang seminimal mungkin namun menghasilkan klaster yang akurat, efektif dan efisien.
3. Coba dilakukan perbandingan dengan *Incremental Genetic K-means Algorithm* (IGKA) yang mungkin dapat menambah tujuan clustering dari tujuan sebelumnya. Dimana IGKA merupakan perkembangan dari FGKA yang idenya mampu menghitung TWCV dan menentukan titik pusat klaster secara *incrementaly* (meningkat teratur) pada kondisi nilai probabilitas mutasinya kecil. Sehingga diperoleh kesimpulan sementara bahwa FGKA baik digunakan untuk nilai probabilitas mutasi yang besar, sedangkan IGKA baik digunakan untuk nilai probabilitas mutasi yang kecil.

Daftar Pustaka

[1] Goharian, Grossman. *Data Preprocessing.* Illinois Institute of Technologi, 2003

[2] Goldberg, David E. *Genetic algorithms in search, optimization, and machine learning.* The University of Alabama, Addison-Wesley Publishing Company,INC.1985

[3] Jiawei Han and Micheline Kamber. *Data Mining : Concepts and Techniques.* Intelligent Database Systems Research Lab, School of Computing Science, Simon Fraser University.

[4] Kusumadewi,Sri.*Artificial Inteligence(Teknik dan Aplikasinya).*2003

[5] Lance D Chambers. *Practical Handbook of Genetic Algorithm Complex Coding System* .Perth, Wester Australia 1998.

[6] Mattison, Rob. *Data Warehousing and Data mining for Telecommunications.* Artech house, INC.1997

[7] Pang-Ning Tan, Michael Steinbach and Vipin Kumar. *Introduction to Data Mining.* University of Minnesota and Army High Performance Computing Research Center

[8] Pramudiono Iko. *Pengantar Data Mining : Menambang Permata Pengetahuan di Gunung Data.* http://www.ilmukomputer.com.2003

[9] Suyanto. *Algoritma Genetika dalam MATLAB.* Andi. Jogjakarta. 2005.

[10] Yi Lu, Shiyong Lu, Farshad Fotouhi, Youping Deng, and Susan Brown. *Incremental Genetic K-means Algorithm and its Application in Gene Expression Data Analysis.* Departement of Computer Science, Wayne State University ,USA.

[11] Yi Lu, Shiyong Lu, Farshad Fotouhi, Youping Deng, and Susan Brown, *FGKA : A Fast Genetic K-means Algorithm,* in proceeding of the 19th ACM Symposium on Applied Computing (SAC), Nicosia, Cyprus, March, 2004

[12] Zengyou He, Xu Xiao Fei, Deng Shengchun and Song Yufu. *dNumber*: *A Fast Algorithm For Very Large Categorical Dataset.* Departemen of computer science and engeneering. Harbin Institute of Technologi. China.