DATA MINING UNTUK KLASTERING DATASET MULTIDIMENSIONAL MENGGUNAKAN ALGORITMA FORGY, ISODATA DAN K-MEANS

David[*)]

Abstract : *At this time, information can be obtained from a pile of very large size of data, hence the need to data mining activities that are still hidden to further processed into useful knowledge in decision making. The process of extraction of information from the collection of data that is stored with the data mining. One of the methods in Data Mining is clustering. Clustering aims to segment the physical or abstract objects in the form of classes or objects similar. Some partitioning clustering are Forgy, ISODATA and K-Means algorithms. On this research analyzed the influence of the changes made on the application of the level of accuracy and the resulting clusters, and system parameters that apply. For the purpose of software is built as a media test Forgy, ISODATA and K-Means algorithms.Level of accuracy can be improved by using threshold, maximum number of iterations and increase the value of the parameter. Comparison is made the test of clustering time process, the variance ratio and the value of Sum Squared Error. Experiments are carried out on the number of datasets.This research concludes that the three clustering methods has similar value of variance ratio, and high intraclass similarity and low interclass similarity. The clustering algorithm using K-Means requires the longest time..*

Keywords*: Clustering, Forgy, ISODATA, K-Means, Cluster Analysis*

Perkembangan teknologi informasi saat ini kian hari kian pesat, sehingga mendorong tingginya kebutuhan informasi. Informasi dapat diperoleh dari suatu timbunan data-data yang berukuran sangat besar. Data tersebut menjadi sumber data yang terbesar dan sangat berharga untuk setiap pengguna karena didalamnya terdapat kumpulan informasi yang saling terhubung. Setiap kali informasi ditambahkan, makin besar ukuran data dan semakin sulit pula untuk mencari informasi yang benar-benar yang diinginkan oleh karena itu dibutuhkan suatu teknologi untuk mendapatkan informasi yang benar-benar diinginkan tersebut.

Informasi yang didapat dari data sangat sedikit yang dapat dimanfaatkan, oleh karena itu perlu adanya aktivitas penggalian (ekstraksi) data yang masih tersembunyi untuk selanjutnya diolah menjadi pengetahuan yang bermanfaat

---

*) *David, S.Kom. M.cs. adalah dosen Sekolah Tinggi Manajemen Informatika dan Komputer Pontianak*

dalam pengambilan keputusan. Proses ekstraksi informasi dari kumpulan data-data yang tersimpan disebut dengan *data mining*.

Turban dan Aronson (1998) mengemukakan bahwa *data mining* merupakan suatu bentuk yang biasa digunakan untuk menggambarkan *knowledge discovery* dari sebuah database, *knowledge extraction, data archaeology, data exploration, data pattern processing, information harvesting* dan *software*. Jadi Data mining adalah suatu proses analisis data dari berbagai perspektif yang berbeda dan memberikan ringkasannya dalam bentuk informasi yang bermanfaat, memungkinkan pemakai untuk melakukan analisis data dari dimensi yang berbeda-beda, dan secara teknis merupakan proses menemukan korelasi atau bentuk dari suatu kumpulan database relasional. Salah satu penerapan teknik *data mining* adalah *clustering*. Banyak sekali metode *clustering* yang dapat diimplementasikan, namun penulis menggunakan algoritma *Forgy, ISODATA* dan *K-Means* untuk klastering.

Data Mining

Data mining merupakan teknologi yang menggabungkan metoda analisis tradisional dengan algoritma yang canggih untuk memproses data dengan volume besar. Data mining adalah suatu proses mengolah data-data yang ada dengan metode dan algoritma tertentu, yang bertujuan mengekstrak pola yang berguna dari data yang besar tersebut (Han dan Micheline, 2001).

Adapun tugas data mining antara lain (Han dan Micheline, 2001): (1) Metoda prediksi, yaitu menggunakan beberapa variable untuk memperkirakan suatu nilai yang tidak diketahui dari variable yang lain. Sasaran pada tugas ini adalah memprediksikan nilai atribut tertentu berdasarkan nilai atribut yang lain. Atribut yang diprediksi dikenal sebagai target atau variabel yang tergantung pada variabel lain, atribut yang digunakan selama membuat prediksi dikenal sebagai penjelasan (*explanatory*) atau variabel yang bebas. Contohnya antara lain *Classification, Regression* dan *Deviation* Detection. (2) Metoda deskripsi, yaitu mencari suatu pola yang dapat ditafsirkan manusia sehingga data dapat digambarkan atau diuraikan. Sasaran pada tugas ini adalah memperoleh pola (kecenderungan korelasi, *cluster* dan anomali) yang menyimpulkan hubungan dalam data. Tugas deskriptif *data mining* memerlukan teknik *postprocessing* untuk validasi dan

kejelasan hasil. Contohnya antara lain *Clustering* dan *Association Rule Discovery*.

Clustering

Pada dasamya clustering terhadap data adalah suatu proses untuk mengelompokkan sekumpulan data tanpa suatu atribut kelas yang telah didefinisikan sebelumnya, berdasarkan pada prinsip konseptual clustering yaitu memaksimalkan dan juga meminimalkan kemiripan intra kelas (Ming dan Hou, 2004). Misalnya, sekumpulan obyek-obyek komoditi pertama-tama dapat di clustering menjadi sebuah himpunan kelas-kelas dan lalu menjadi sebuah himpunan aturan-aturan yang dapat diturunkan berdasarkan suatu klasifikasi tertentu.

Proses untuk mengelompokkan secara fisik atau abstrak obyek-obyek ke dalam bentuk kelas-kelas atau obyek-obyek yang serupa, disebut dengan *clustering* atau *unsupervised classification*. Melakukan analisa dengan *clustering*, akan sangat membantu untuk membentuk partisi-partisi yang berguna terhadap sejumlah besar himpunan obyek dengan didasarkan pada prinsip *divide and conquer* yang mendekomposisikan suatu sistem skala besar, menjadi komponen-komponen yang lebih kecil, untuk menyederhanakan proses desain dan implementasi.

Pada dasarnya terdapat dua tipe clustering, yaitu: 1) *Partitional Clustering*: Tipe cluster yang benar-benar terpisah antara sekelompok obyek dengan sekelompok obyek lainnya dan 2) *Hierarchical clustering* : Sekelompok cluster yang terorganisasi sebagai suatu pohon hirarki (*hierarchical tree*). **Algoritma *Forgy* dan *ISODATA***

Gose dkk. (1996) menyatakan bahwa Algoritma *Forgy* merupakan salah satu metode klastering yang sederhana. Disamping menggunakan data, yang menjadi inputan pada algoritma adalah *k*, yaitu jumlah klaster yang akan dibentuk. Sampel *k* disebut dengan *seed point*. *Seed point* dipilih secara acak untuk membantu pemilihan klaster.

Gose dkk. (1996) menyatakan bahwa algoritma *ISODATA* (*Iterative Self-Organizing Data Analysis Techniques*) merupakan pengembangan dari algoritma *Forgy* dan *K-means*. Sama seperti kedua algoritma tersebut, algoritma *ISODATA* mencoba meminimalkan kuadrat kesalahan (*squared error*).

Algoritma K-Means

K-Means Cluster merupakan algoritma *clustering* yang berulang–ulang. Algoritma *K-Means* dimulai dengan pemilihan secara acak *K* untuk *cluster centroid* (nilai *K* umumnya ditetapkan dahulu). Setiap kejadian membentuk sebuah klaster, dari banyaknya *cluster* dicari sebagai *center*. Jika jumlah anggota klaster sama dengan nilai *K* maka klaster tersebut ditutup. Selanjutnya setiap kejadian yang telah terbentuk *centroid* akan diproses ulang. Proses ini akan diulang sampai *cluster centroid* menjadi stabil.

Analisa Klaster *(Cluster Analysis)*

Analisa cluster adalah suatu teknik analisa *multivariate* (banyak variabel) untuk mencari dan mengorganisir informasi tentang variabel tersebut sehingga secara relatif dapat dikelompokkan dalam bentuk yang homogen dalam sebuah klaster (Nadler dan Smith, 1993). Analisis cluster diukur dengan menggunakan nilai *variance* atau *error ratio*. *Variance* digunakan untuk mengukur nilai penyebaran dari data-data hasil clustering dan dipakai untuk data bertipe *unsupervised*. Secara umum, bisa dikatakan sebagai proses menganalisa baik tidaknya suatu proses pembentukan *cluster*. Analisa cluster bisa diperoleh dari kepadatan cluster yang dibentuk (*cluster density*). Kepadatan suatu cluster bisa ditentukan dengan *variance within cluster* (V_w) dan *variance between cluster* (V_b). Varian tiap tahap pembentukan cluster.

Salah satu metode yang digunakan untuk menentukan cluster yang ideal adalah batasan *variance*, yaitu dengan menghitung kepadatan *cluster* berupa *variance within cluster* (V_w) dan *variance between cluster* (V_b) (Veenman dkk, 2002). Cluster yang ideal mempunyai V_w minimum yang merepresentasikan *internal homogenity* dan maksimum V_b yang menyatakan *external homogenity*. Cluster disebut ideal jika memiliki nilai V_w seminimal mungkin dan V_b semaksimal mungkin.

Meskipun minimum V_w menunjukkan nilai cluster yang ideal, tetapi pada beberapa kasus kita tidak bisa menggunakannya secara langsung untuk mencapai global optimum. Jika dipaksakan, maka solusi yang dihasilkan akan jatuh pada local optima.

Renals (2009) menyatakan bahwa metode lainnya untuk menganalisis hasil klaster adalah dengan menghitung *Sum Squared Error (SSE)*. Untuk setiap *data point*, nilai kesalahan didapatkan dari perhitungan jarak dengan klaster

terdekatnya. Untuk mendapatkan nilai SSE, nilai error yang dikuadratkan kemudian dijumlahkan semua. Perhitungan jarak digunakan persamaan *squared Euclidean distance*. (Gose, dkk, 1996 dan Renals, 2009).

Salah satu cara untuk mereduksi *SSE* adalah dengan meningkatkan nilai K (jumlah klaster). Hasil klastering yang baik yaitu memiliki nilai *SSE* dengan *error* terkecil. Klastering yang baik dengan K yang lebih kecil memiliki nilai *SSE* yang rendah daripada klastering dengan K yang besar.

Metode Penelitian

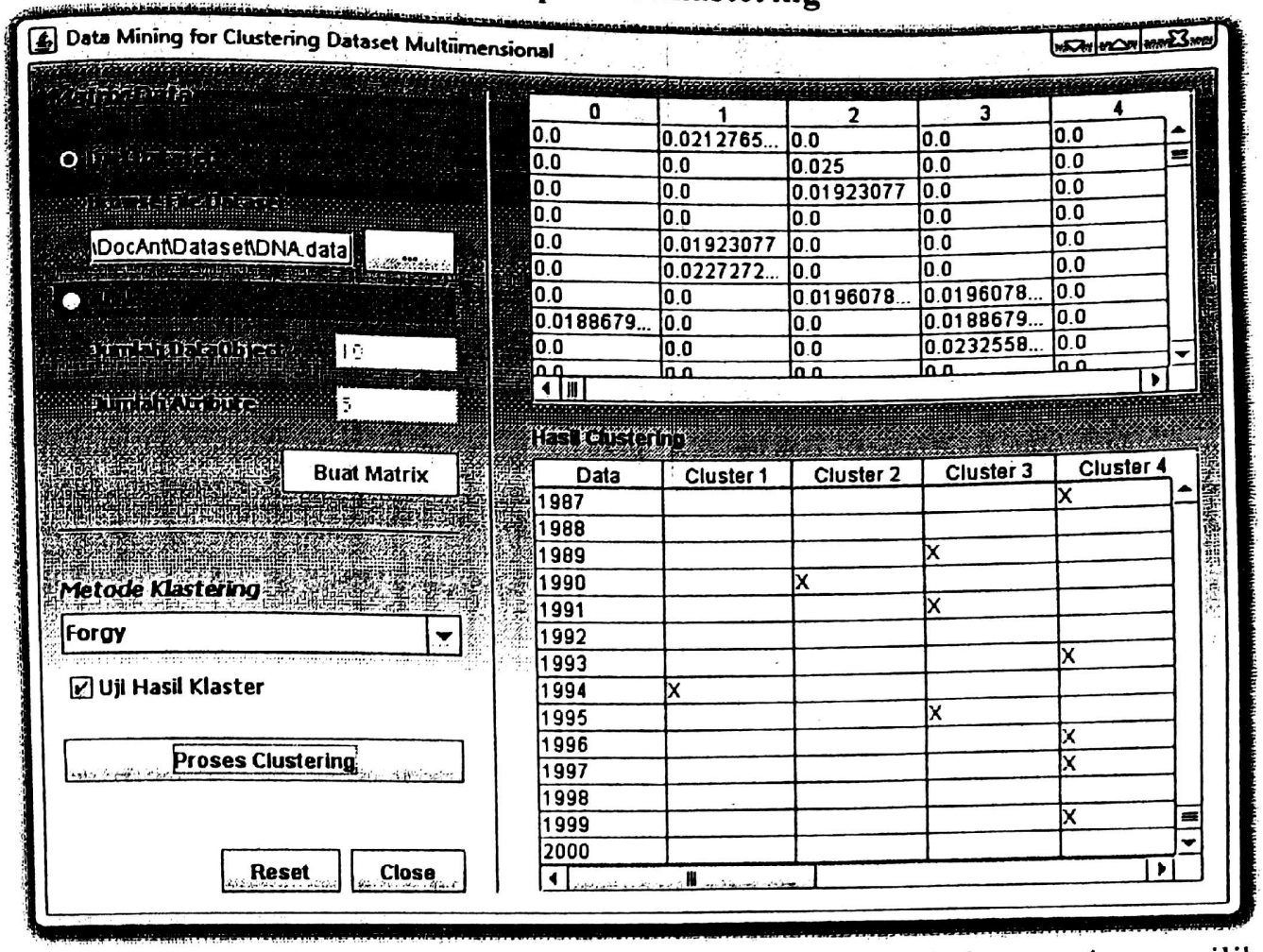
Dalam penelitian ini yang dijadikan obyek penelitian adalah kumpulan Dataset yang dapat diunduh di http://www.cs.sfu.ca/~wangk/ucidata/dataset/. Dalam mengumpulkan data-data penulis menggunakan metode: 1) Dokumentasi, yaitu mengambil dataset yang dibutuhkan sebagai data pengujian, 2) Studi Literatur, yaitu mempelajari semua literatur yang berkaitan dengan klastering menggunakan *Forgy, ISODATA dan K-Means*.

Penelitian ini menggunakan metode pengembangan perangkat lunak yang terdiri dari tahap-tahap sebagai berikut: 1) Analisa, Merupakan tahap untuk menganalisa data yang diperoleh sebagai bahan pengujian, 2) Perancangan. Merupakan tahap perancangan terhadap aplikasi klastering, 3) Implementasi. Merupakan tahap untuk mengimplementasikan hasil rancangan sistem menjadi sebuah perangkat lunak, 4) Pengujian. Merupakan tahap uji coba dari perangkat lunak yang telah diimplementasikan, dan 5) Evaluasi dan Perbaikan Kesalahan. Merupakan tahap untuk melakukan evaluasi dan dan perbaikan terhadap kesalahan-kesalahan yang terjadi dalam perangkat lunak yang dibuat.

HASIL

Hasil implementasi dari aplikasi data mining untuk klastering dataset muldimensional dapat dilihat pada gambar 1.

Gambar 1: Hasil Implementasi Aplikasi Klastering



Pada implementasi di atas dapat dilihat bahwa pemakai dapat saja memilih dataset ataupun menggunakan data acak, memilih metode klastering dan uji klaster dalam aplikasi klastering dataset multidimensional.

Pengujian meliputi pengujian nilai *variance*, pengujian nilai *Sum Squared Error*, dan pengujian kinerja aplikasi berdasarkan waktu pemrosesan. Semua eksperiman dilakukan pada sebuah Notebook HP Compaq, INTEL Core 2 DUO T5300, 1.73 GHz, Memory 2 GB dan Sistem Operasi Windows Vista. Tabel 1 menunjukkan sejumlah dataset multidimensional yang digunakan untuk pengujian.

Tabel 1: Tabel Dataset

No	Dataset	Jumlah Record	Jumlah Atribut
1	Wine	178	13
2	Heart	270	13
3	Sonar	208	60
4	Diabetes	768	8
5	DNA	2000	180
6	Waveform	5000	21

7	Page-blocks	5473	10
8	Ann-thyroid	7200	21
9	Letter-recognition	20000	16
10	Shuttle	43500	9

Pengujian klastering dataset tersebut menggunakan algoritma *Forgy* menggunakan parameter sebagai berikut : jumlah klaster (K) = 10 dan maksimum jumlah iterasi = 1000. Algoritma ISODATA dengan parameter sebagai berikut : jumlah cluster (K) = 5, maksimum jumlah Iterasi = 1000, *minimum number threshold* =2, *standard deviation threshold* = 0.5, *minimum distance threshold* = 0.5 dan maksimum *number threshold* = 1. Algoritma K-Means dengan parameter sebagai berikut : jumlah klaster (K) = 10 dan maksimum jumlah iterasi = 1000. Tabel hasil pengujian sejumlah dataset dapat dilihat pada tabel 2.

Berdasarkan hasil pengujian klastering dataset pada tabel 2, dapat dilihat bahwa klastering menggunakan algoritma Forgy dan K-Means selalu menghasilkan jumlah klaster yang tetap, sedangkan untuk klastering menggunakan algoritma ISODATA menghasilkan jumlah klaster yang tidak tetap, hal ini dikarenakan pada algoritma ISODATA menggunakan konsep *Merge* dan *Split*.

Tabel 2: Tabel Hasil Pengujian Klastering

No	Dataset	Algoritma Klastering	Waktu	Rasio Variance	SSE	Klaster yang terbentuk
1	Wine	Forgy	00m:01s	0,655120700	0.138248950	10
		ISODATA	00m:01s	0,127443620	0,154697600	3
		K-Means	00m.12s	0,655120730	0,136722790	10
2	Heart	Forgy	00m:01s	0,226439820	0,270828900	10
		ISODATA	00m:02s	0,254021320	0,248047860	2
		K-Means	00m:49s	0,193491090	0,265862580	10
3	Sonar	Forgy	00m:06s	0,258788780	0,624489000	10
		ISODATA	00m:09s	0,284488400	0,698372000	10
		K-Means	00m:37s	0,258788780	0,624489000	10
4	Diabetes	Forgy	00m:12s	0,017448940	0,599548640	10
		ISODATA	00m:15s	0,017407300	0,599150400	7
		K-Means	00m:27s	0,017448938	0,609265400	10

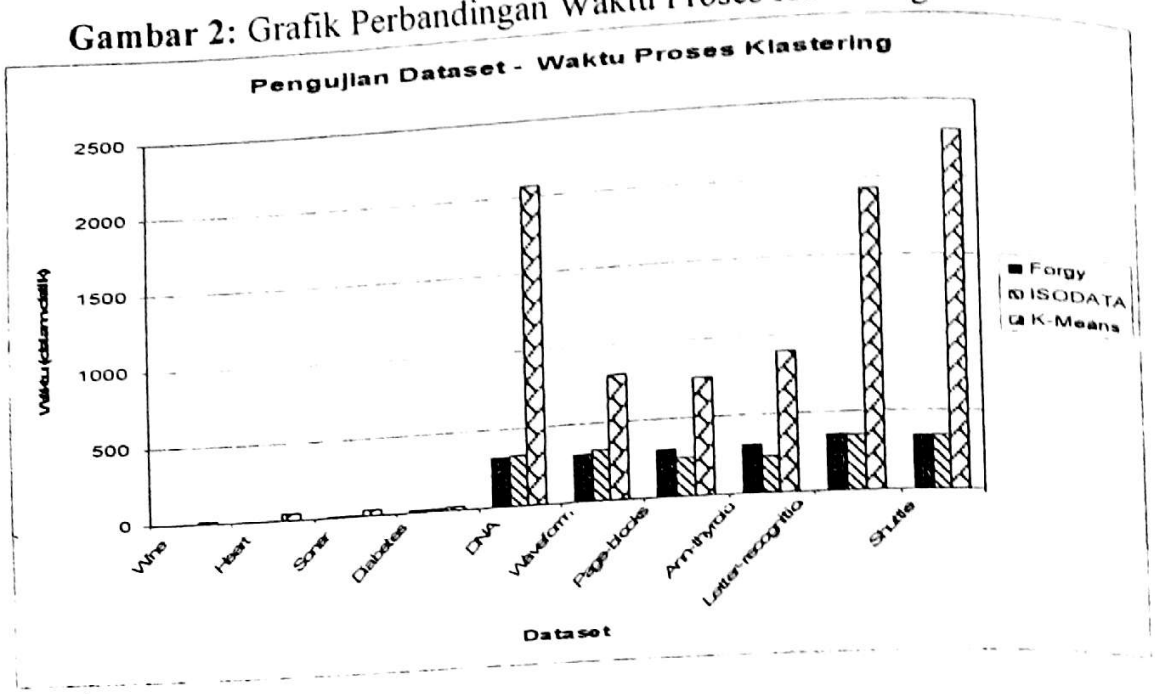
5	DNA	Forgy	05m:19s	0,510946400	1,255185100	10
		ISODATA	05m:25s	0,515618740	1,250335800	3
		K-Means	34m:34s	0,603577500	0,159417510	10
6	Waveform	Forgy	05m:03s	0,140364980	0,764657300	10
		ISODATA	05m:29s	0,150361230	0,947715340	2
		K-Means	13m:21s	0,382668900	0,105100766	10
7	Page-blocks	Forgy	05m:06s	0,535175200	0,542710960	10
		ISODATA	04m:13s	0,599462100	0,581991600	7
		K-Means	12m:41s	0,798548360	0,081877000	10
8	Ann-thyroid	Forgy	05m:12s	0,313329080	0,652628300	10
		ISODATA	03m:55s	0,313329080	0,915112730	10
		K-Means	14m:58s	0,582832800	0,097563290	10
9	Letter-recognition	Forgy	06m:01s	0,217245000	0,515340200	10
		ISODATA	05m:50s	0,193177780	0,502222060	10
		K-Means	32m:27s	0,231225740	0,155977040	10
10	Shuttle	Forgy	05m:42s	0,230935500	0,677955100	10
		ISODATA	05m:47s	0,230778420	0,689446570	9
		K-Means	38m:28s	0,243475700	0,175795020	10

Keterangan : m = menit (*minute*); s = detik (*second*)

Berdasarkan tabel 2 dapat digambarkan grafik pengujian kinerja system berdasarkan waktu klasteringnya. Gambar 2 menunjukkan grafik hasil klastering dengan waktu klastering. Dari grafik tersebut dapat disimpulkan bahwa algoritma *K-Means* memerlukan waktu yang relative lama seiring dengan ukuran multidimensional dari dataset dibandingkan dengan algoritma Forgy dan ISODATA.

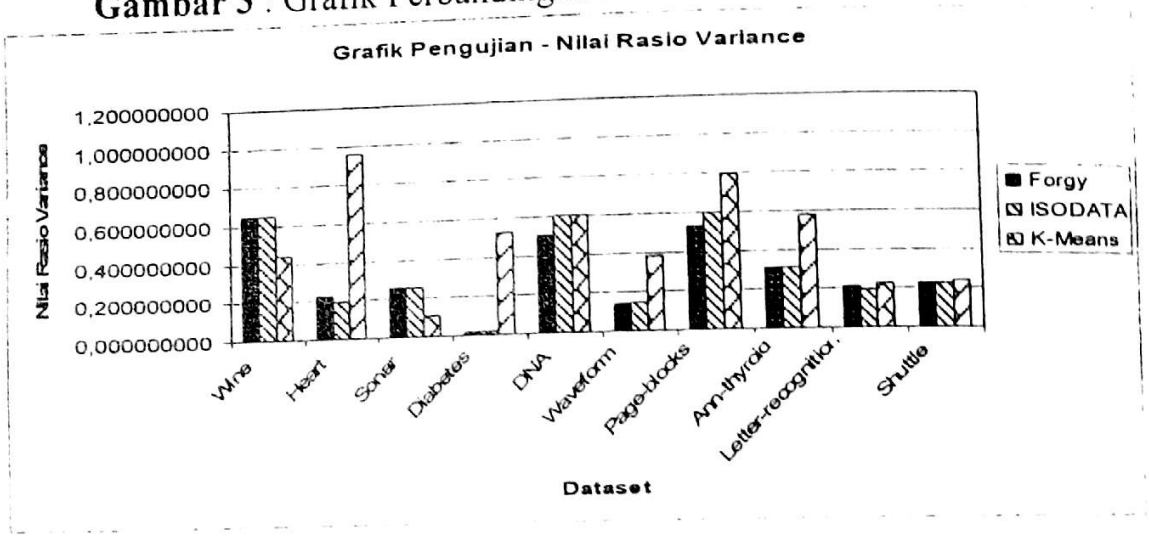
Berdasarkan tabel 2 dapat digambarkan grafik perbandingan nilai *variance* ideal dari ketiga metode klastering tersebut. Gambar 3 memperlihatkan grafik perbandingan nilai rasio *variance*. Dari gambar grafik tersebut dapat dilihat bahwa pada dataset Letter-Recognition dan Shuttle, ketiga algoritma tersebut memiliki nilai rasio *variance* yang hampir sama. Namun secara keseluruhan dapat

Gambar 2: Grafik Perbandingan Waktu Proses Klastering



disimpulkan bahwa algoritma *K-Means* cenderung memiliki nilai variance yang lebih besar daripada kedua metode lainnya, kecuali pada kasus dataset Wine dan Sonar. Pada dataset Sonar, algoritma *Forgy* dan *ISODATA* memiliki nilai variance yang tinggi dibandingkan dengan algoritma *K-Means*.

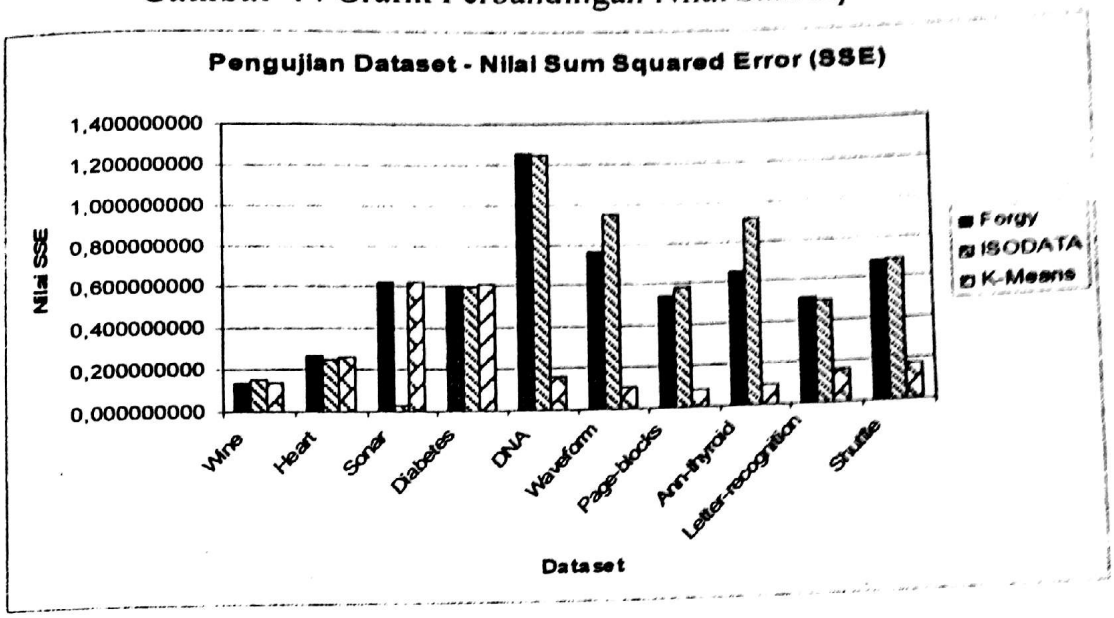
Gambar 3 : Grafik Perbandingan Nilai Rasio *Variance*



Berdasarkan tabel 2 dapat digambarkan grafik perbandingan nilai *Sum Squared Error* (SSE) dari ketiga metode klastering. Gambar 4 menunjukkan grafik perbandingan nilai *SSE*. Dari gambar grafik tersebut dapat disimpulkan bahwa algoritma *K-Means* memiliki nilai *SSE* yang rendah dan dapat dikatakan hasil klasteringnya merupakan klaster yang baik. Nilai SSE antara Algoritma Forgy dan ISODATA memiliki nilai yang hampir sama. Pada dataset Heart,

Sonar, Diabetes dan DNA, nilai SSE menggunakan Algoritma Forgy lebih tinggi dibandingkan dengan algoritma ISODATA.

Gambar 4 : Grafik Perbandingan Nilai *Sum SquareError*



Kesimpulan

Dari hasil penelitian dan pengujian yang dilakukan pada sistem aplikasi klastering dataset menggunakan algoritma *Forgy, ISODATA* dan *K-Means* dapat ditarik kesimpulan sebagai berikut:

a. Sistem klastering menggunakan algoritma *ISODATA* menghasilkan klaster yang tidak pasti. Sedangkan algoritma *Forgy* dan *K-Means* selalu menghasilkan klaster yang tetap.
b. Klastering menggunakan Algoritma *K-Means* memerlukan waktu pemrosesan yang sangat lama, hal ini dipengaruhi oleh perubahan titik pusat klaster dan jumlah iterasi yang diberikan terhadapnya.
c. Secara keseluruhan hasil klastering dari ketiga metode tersebut memiliki nilai variance yang hampir sama dimana masing-masing hasil klastering menghasilkan *intraclass similarity* (kesamaan di dalam klas) yang tinggi dan *interclass similarity* (kesamaan antar klas) yang rendah, hal ini didasarkan pada jarak antar vektor data dengan pusat klaster yang dihasilkan.
d. Hasil klastering menggunakan *K-Means* merupakan hasil klaster yang lebih baik dibandingkan dengan klastering menggunakan *ISODATA* dan *Forgy* yang dapat dilihat dari perolehan nilai *Sum Squared Error* dari sejumlah pengujian dataset.

DAFTAR PUSTAKA

Han. J., dan Kamber. M., 2001, *Data Minings : Concept and Techniques*, Morgan Kaufmann Publishers, United States of Amerika

Ball, G.H. dan Hall, D.J., 1965, *ISODATA: A novel methods of data analysis and pattern classification*, Technical Report AD0699616, Stanford Research Institute, Stanford, CA, U.S., April 1965.

Gose, E., Johnsonbaugh, R., dan Jost, S., 1996, *Pattern recognition and Image Analysis*, Prentice Hall, USA.

Memarsadeghi, N., Mount, D.M., Netanyahu, N.S., dan Moigne, J.L., 2007, A Fast Implementation of the ISODATA Clustering Algorithm, *International Journal of Computational Geometry and Applications*, 17(1):71-103, February 2007.

Ming, W.H., dan Hou. C.J., 2004, *Cluster analysis and visualization*, Workshop on Statistics and Machine Learning, Institute of Statistical Science, Academia Sinica.

Nadler, M dan Smith, E.P., 1993, *Pattern recognition Engineering*, John Wiley & Sons.,Inc., USA.

Renals, S., 2009, ***Clustering***, Learning and Data Note 3 (v2.2).

Turban, E. dan Aronson, J, 1998, *Decision Support Systems and Intelligent Systems*, Fifth Edition, Penerbit Prentice-Hall, Inc.

Veenman, C.J.., Reinders, M.J.T., dan Backer, E.., 2002, A maximum variance cluster algorithm, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1273-1280.

Weisstein, E.W., 2008, *K-Means Clustering Algorithm*, MathWorld-A Wolfram Web Resourcem http://mathworld.wolfram.com/K-MeansClusteringAlgorithm.html