OPTIMASI PUSAT CLUSTER K-PROTOTYPE DENGAN ALGORITMA GENETIKA

Pivin Suwirmayanti [1], I Ketut Gede Darma Putra [2], I Nyoman Satya Kumara[3]
[1] Mahasiswa Magister Teknik Elektro, Program Pasca Sarjana Universitas Udayana
[2, 3] Staf Pengajar Magister Teknik ElektroProgram Pasca Sarjana Universitas Udayana
Kampus Sudirman Denpasar
Email : pivin.hena@yahoo.co.id

Abstrak

Teknik *cluster*ing saat ini telah banyak digunakan untuk mengatasi permasalahan yang terkait dengan segementasi data. Implementasi *cluster*ing ini dapat diterapkan pada berbagai bidang sebagai contoh dalam hal pemasaran, *cluster*ing dapat digunakan sebagai metode untuk mengelompokkan data. Metode Clustering memiliki tujuan untuk mengelompokkan beberapa data ke dalam beberapa kelompok data sehingga kelompok yang terbentuk memiliki kemiripan data, secara umum proses clustering diolah menggunakan tipe data numerik, namun pada kenyataannya proses pengelompokan data tidak hanya menggunakan tipe data numerik, terdapat juga tipe data kategorikal. Untuk itu penulis menggunakan metode *K-Prototype* yang dioptimasi dengan Algortima Genetika dimana data uji yang digunakan adalah Data German Credit yang memiliki tipe data numerikal dan kategorikal. Dalam penelitian dilakukan perbandingan kinerja antara metode *K-Prototype* dengan Algoritma Genetika, dengan metode *K-Prototype* Tanpa Algortima Genetika, dan metode *K-Means*. Dari beberapa hasil percobaan yang dilakukan metode *K-Prototype* dengan Algoritma Genetika menghasilkan hasil yang terbaik dari metode *K-Prototype* tanpa Algortima Genetika, dan metode *K-Means*

Kata Kunci : *Clustering, Optimasi, K-Prototype, Algoritma Genetika*

1. PENDAHULUAN

Analisa *Cluster* saat ini merupakan metode yang banyak digunakan dalam pengelompokan data menjadi segmen–segmen yang lebih kecil. Analisa *Cluster* merupakan suatu teknik yang digunakan untuk membagi sekumpulan obyek ke dalam k kelompok sehingga nilai dalam setiap kelompok adalah homogen dengan mengacu kepada atribut tertentu berdasarkan kriteria tertentu [1]. Kesamaan nilai (*homogenity*) dalam setiap segmen yang diwakili oleh *cluster* menggambarkan kesamaan pola perilaku pelanggan[1]. Dengan analisa *cluster* ini, kesamaan pola prilaku pelanggan dapat digali melalui *cluster-cluster* yang terbentuk. Semakin akurat *cluster* yang terbentuk maka akan semakin jelas kesamaan pola prilaku dari pelanggan. Sehingga dengan demikian, para pelaku bisnis dapat menentukan strategi pemasaran dengan lebih akurat, berdasarkan pada pola prilaku pelanggan yang didapatkan dari proses analisa *cluster* tersebut.

Teknik c*lustering* saat ini juga telah banyak digunakan untuk mengatasi permasalahan yang terkait dengan segementasi data. Implementasi *clustering* ini dapat diterapkan pada berbagai bidang sebagai contoh dalam hal pemasaran, *clustering* dapat digunakan sebagai metode untuk mengelompokkan pelanggan yang memiliki kesamaan dalam perilaku belanja. Metode *clustering* memiliki tujuan untuk mengelompokkan beberapa data ke dalam beberapa kelompok data sehingga kelompok yang terbentuk memiliki kemiripan data. Proses yang dilakukan pada saat *clustering* adalah menempatkan obyek yang memiliki kemiripanatau memiliki jarak yang terdekat dalam satu*cluster*.Dengan demikian sebuah objek akan memiliki kemiripan yang sama dengan objek lain dalam satu cluster dan memiliki kemiripan yang berbeda dengan satu cluster lainnya.*Clustering*sering disebut dengan teknik *unsupervised learning* dimana tidak memerlukan fase training.

Berkembangnya penelitian tentang teknik *clustering*, telah ditemukan berbagai algoritma yang bisa menghasilkan *cluster* dengan tingkat akurasi yang semakin baik. Salah satu teknik yang menggunakan Algoritma Genetika yang dikombinasikan dengan metode *K-MeansClustering*, untuk mendapatkan jumlah *cluster* yang optimal [2]. Kombinasi Metode GA dan *K-Means* juga dimanfaatkan untuk pengelompokan pelanggan dalam membuat *recomender system* pada *online* shoping market [3]. Dari hasil penelitian tersebut, bisa disimpulkan bahwa GA *K-Means* mampu menghasilkan pengelompokan (*clustering*) yang lebih baik dibandingkan dengan *Self Organising Map* (SOM) yang berbasis *neural network*. Penelitian terkait untuk mengoptimalkan Algoritma *K-Means* dalam menentukan pusat awal *cluster*, dimana hasil penelitian menunjukkan bahwa algoritma *K-Means* memiliki kelemahan tidak hanya memiliki ketergantungan pada data awal, tetapi juga konvergensi yang cepat dan kualitas *clustering*[4]. Untuk memperoleh *cluster* yang efektif dan akurat, maka Min Feng dan Zhenyan-wang mengoptimalkan Algoritma *K-Means* (PKM) dengan Algoritma Genetika menjadi sebuah Algoritma Hibrid (PGKM).

Percobaan menunjukkan bahwa algoritma itu memiliki kualitas *cluster* dan performance yang baik [4].

Publikasi menggunakan algoritma genetik (*Genetic Algorithm*, GA) dan dikombinasikan dengan metode *clustering* yang sangat populer, yaitu *K-Means* untuk menemukan variabel yang valid dan jumlah *cluster* optimal secara simultan[5]. Hasil penelitian menunjukkan, gabungan algoritma genetik dan *K-Means* berhasil menghilangkan variabel yang tidak relevan dan menghasilkan jumlah *cluster* secara otomatis, dan berhasil meningkatkan hasil pengelompokan pelanggan secara signifikan. Penelitian terkait yang memperbaiki *K-MeansClustering* dengan menggunakan Algoritma Genetik, dimana disebutkan *ClusteringK-Means* adalah sebuah algoritma *clustering* yang populer berdasarkan hasil partisi data [6].

Untuk meningkatkan kualitas dari *clusterK-Means* menggunakan Algoritma Genetika dan hasil penelitian menunjukkan bahwa algoritma yang diajukan mencapai hasil yang lebih baik. Penelitian yang menyebutkan ada sebuah algoritma *clustering* baru yang diusulkan yaitu *Modified Genetic Algorithm Initializing K-Means* (MGAIK) [7]. MGAIK diinspirasi oleh sebuah metode *initialization* algoritma genetik untuk *K-Meansclustering* tapi beberapa fitur perbaikan dari GAIK. Akhirnya, ketika perbandingan yang dilakukan mendapatkan hasil bahwa MGAIK lebih baik dari yang sederhana Algoritma Genetika .

Keseluruhan penelitian tersebut di atas, belum ada penelitian yang memperhatikan kemungkinan munculnya kesalahan pada data karena adanya perbedaan tipe data dalam dataset yang digunakan. Misalnya dalam penelitian yang dilakukan oleh Liu dkk, data uji yang digunakan memiliki tipe data campuran yaitu data dengan tipe numerik dan kategorikal. Sementara itu penelitian yang dilakukan menunjukkan bahwa, penanganan data kategorikal dengan metode untuk data numerik tidak selalu memberikan hasil yang berguna, karena tidak semua data *categorical* di dunia nyata disajikan dalam bentuk terurut(*ordered*) [1].

Penelitian dengan algoritma yang disebut dengan *K-Prototype*, untuk menangani *clustering* data tipe campuran numerik dan kategorikal. Penelitian ini menyajikan Algoritma *K-Prototype* untuk *cluster* data set yang besar yang ada di dunia nyata. *K-Prototype* adalah salah satu metode *clustering* yang berbasis partitioning. Algoritma ini adalah hasil pengembangan antara *K-Means* dan *K-ModesClustering* untuk menangani data campuran bertipe numerik dan kategorikal. K-Protoype memiliki keunggulan karena algoritmanya yang tidak terlalu kompleks dan mampu menangani data yang besar lebih baik dibandingkan dengan algoritma yang berbasis hierarki [8].

Hasil yang diperoleh dari optimal dengan metode *ClusteringK-Prototype* akan dioptimasi dengan Algoritma Genetika.Pada Algoritma Genetika dapat digunakan dalammenyelesaikan masalah optimasi yang lebih sukar dan kompleks[9].Algoritma Genetika merupakan algoritma optimalisasi dan pencarian yang didasarkan pada prinsip genetika dan seleksi natural. Algoritma Genetika sebenarnya terinspirasi dari prinsip genetika dan seleksi alam (teori Darwin) yang ditemukan pertama kali oleh John Holland dari Universitas Michigan, Amerika Serikat, melalui sebuah penelitian yang dipopulerkan oleh David Goldberg.

Berdasarkan pada fakta-fakta yang didapatkan dari beberapa penelitian sebelumnya maka penulis melalui penelitian ini menggunakan metode *K-Prototype* yang dioptimasi dengan Algoritma Genetika untuk melakukan optimasi pada pusat *cluster*, untuk mendapatkan pusat *cluster* yang optimal, sehingga akurasi dari *cluster* menjadi lebih baik dimana data uji yang aka digunakan adalah German Credit Dataset. German Credit Dataset adalah dataset yang banyak digunakan dalam permasalahan klasifikasi untuk penilaian kelayakan pemberian kredit.

2. METODELOGI

2.1 Gambaran Umum Sistem

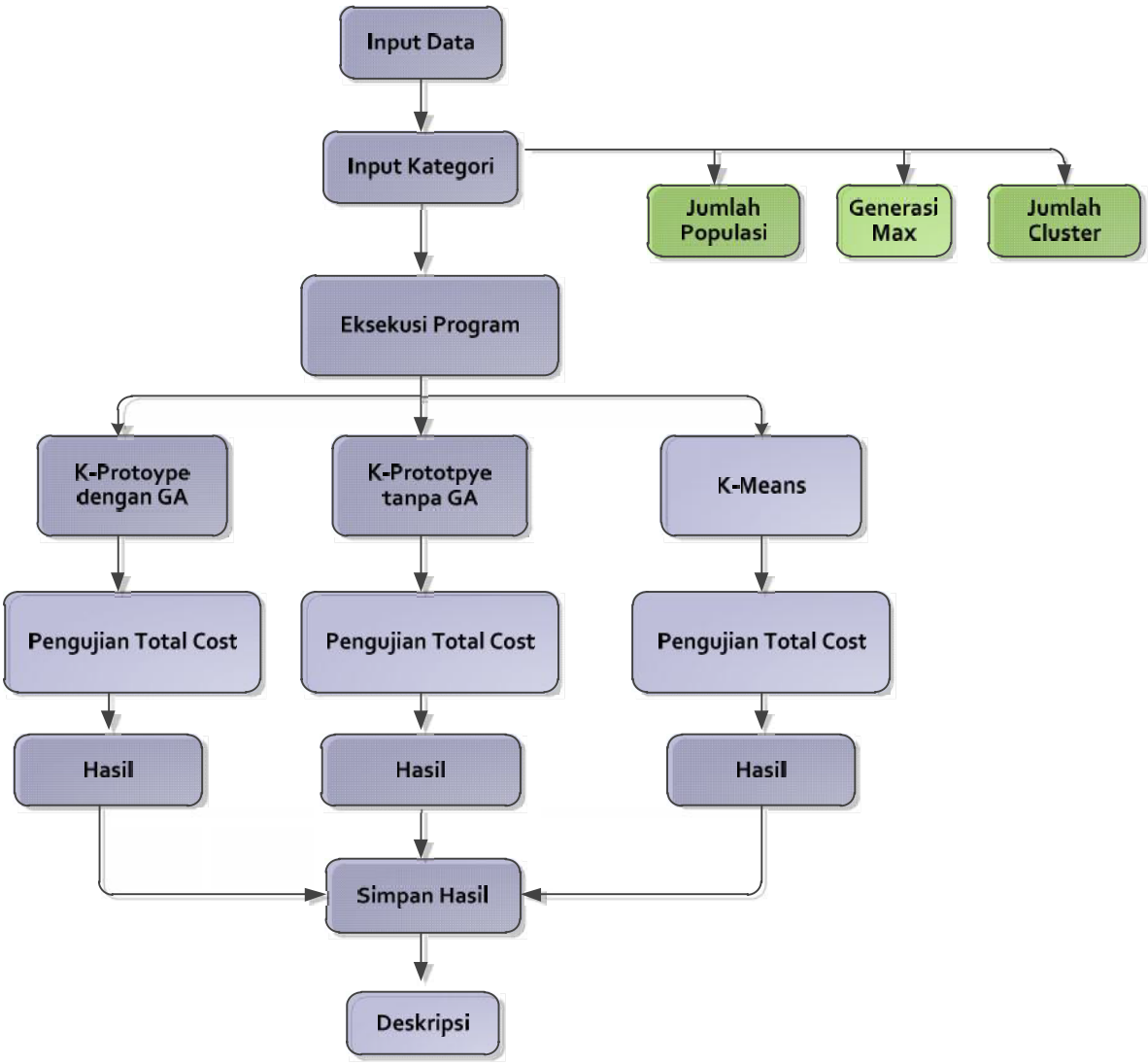
Gambaran umum sistem yang dibuat meliputi beberapa proses yaitu input data uji, untuk input kategori ada 3 hal yang mempengaruhi yaitu Jumlah Populasi bertujuan untuk menentukan jumlah dari individu yang digunakan dalam proses komputasi dalam satu siklus proses evolusi, Generasi Max untuk menentukan berapa kali proses iterasi yang diperlukan, dan jumlah *cluster* untuk menentukan jumlah kelompok data yang dihasilkan. Selanjutnya ada eksekusi program untuk metode yang digunakan yaitu *ClusteringK-Prototype*dengan Algoritma Genetika dan metode penguji yang digunakan yaitu *K-Prototype* dan *K-Means*. Untuk pengukurannya sama-sama menggunkana total jarak dengan menggunkana *Cost function Criterion*.

2.3 Data Uji

Penelitian ini menggunakan data uji yang banyak dipakai dalam permasalahan klasifikasi untuk penilaian kelayakan pemberian kredit (*German Credit Dataset)*, data set ini didonasikan oleh Prof. Hofman dari Hamburg University, Jerman. *Dataset* ini terdiri dari 1000 record dan 20 variabel ditambah dengan sebuah variabel target atau variabel response, dimana 13 variabel diantaranya bertipe kategori dan sisanya sebanyak 7 variabel bertipe numerik. *German Credit Dataset* ini dapat diunduh di UCI *Machine Learning Repository.*

Studi kasus penelitian ini mengenai segmentasi pasar, proses dilakukan dengan mengelompokkan pasar menjadi lebih homogen, yang sebelumnya memiliki prilaku yang heterogen, sehingga membantu para pelaku bisnis dalam mementukan strategi pemasaran.

..



Gambar 1. Gambaran Umum Sistem

.

C. Metode *Clustering K-Prototype* dengan Algoritma Genetika

*K-Prototype*adalah salah satu metode *clustering* yang berbasis *partitioning*. Algoritma ini menangani *clustering* pada data dengan campuran atribut bertipe numerik dan kategorikal. Pada proses dengan *K-Prototype*disini dilakukan beberapa proses yaitu :

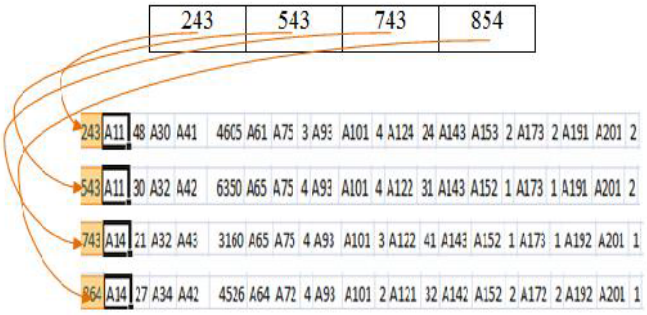
1. Inisialisasi Populasi Awal

Fase inisialisasi populasi ini digunakan untuk menentukan sejumlah kromosom awal yang akan digunakan untuk komputasi selanjutnya. Panjang gen ini adalah sama dengan jumlah K (jumlah cluster) dari proses clustering. Dimana masing-masing nilai yang ada pada gen akan mewakili nomorrecord data pada proses clustering. Nilai yang terkandung dalam gen berupa integer yang dibangkitkan secara acak dengan nilai 0-1000 sesuai jumlah record data.

2. Clustering menggunakan K-Prototype

Setiap kromosom yang terbentuk pada fase inisialiasi populasi akan melewati proses *Clustering*. Jumlah *cluster* akan ditentukan oleh nilai gen terakhir yang ada pada setiap kromosom. Algoritma yang digunakan untuk melakukan *clustering* pada data dengan tipe campuran numerik dan kategorikal ini adalah *K-Prototype*. Algoritma ini dipilih karena sangat sederhana dari sisi kompleksitas algoritma dan mampu menangani data dengan ukuran yang sangat besar dan dengan tipe data campuran. Pada proses c*lustering* dengan *K-Prototype*disini dilakukan beberapa langkah yang dilakukan.

Langkah pertama adalah inisialisasi awal *prototype*menggunakan urutan data record yang sesuai dengan nilai yang ada pada gen di masing-masing kromosom. Misalnya:



Gambar 2. Inisialisasi Awal *Prototype*

Langkah kedua adalah melakukan pengukuran jarak objek ke semua *prototype* dan tempatkan objek pada *cluster* terdekat. Alokasi objek di dalam *X* ke *Cluster* dengan *prototype*terdekat. Tahap ini algoritma *K-Prototype*mengalokasikan semua objek didalam *dataset* ke *cluster* dimana *prototype* dari *cluster* tersebut memiliki jarak yang paling dekat ke objek data. Langkah ketiga adalah Realokasi objekjika terjadi perubahan *prototype.* Setelah semua objek dalam *X* selesai dialokasikan, selanjutnya akan dilakukan pengukuran ulang jarak antara semua objek di dalam *X* terhadap semua *prototype*yang ada. Proses ini akan terus dilakukan sampai tidak ada lagi perubahan *prototype.*

3. Evaluasi Fitness Menggunakan Cost Function Criterion

Dalam penelitian ini, *Clustering Criterion* yang digunakan adalah *cost function*, atau biaya yang dihabiskan untuk menempatkan objek pada *cluster* yang bersesuaian. Semakin kecil biaya *(cost)* yang dikeluarkan, maka semakin bagus hasil *clustering* yang diperoleh.Hal ini berbanding terbalik dengan fungsi *fitness* yang diharapkan, dimana fungsi *fitness* mengharapkan nilai yang semakin besar dari hasil evaluasi kromosom.Untuk mencari nilai fitness dilakukan pengukuran hasil c*lustering*dengan menggunakan *Clustering Criterion. Clustering Criterion* inilah yang akan dijadikan nilai fitness dari setiap kromosom yang telah dievaluasi. Semakin tinggi nilai fitness, maka semakin bagus kromosom yang dievaluasi tersebut. *Clustering Criterion* yang digunakan adalah *cost function*, atau biaya yang dihabiskan untuk menempatkan objek pada *cluster* yang bersesuaian. Solusi umum yang digunakan untuk menentukan apakah bagus atau tidaknya suatu partisi adalah dengan menggunakan *Clustering Criterion.* Dalam hal ini, *Clustering Criterion* yang digunakan adalah dengan mencari *cost function* seperti yang digambarkan seperti dibawah ini.

Cost function yang dipergunakan secara luas adalah penelusuran terhadap matriks dispersi dalam *cluster* (*within cluster dispersion matrix*). Salah satu cara untuk mendefinisikan *cost function* ini adalah dengan formula berikut[10].

$$E = \sum_{l-1}^{k}\sum_{i-1}^{n} y_{il} d(X_i, Q_l) \quad \text{.....(1)}$$

Disini, Ql=[ql1,ql2, ... ,qlm] adalah vector representatif atau sering disebut dengan *prototype* untuk *cluster*, dan y_{il} adalah sebuah elemen dari matrik partisi Y_{nxl}. *d* adalah pengukuran kesamaan (similarity measure) yang pada umumnya menggunakan jarak Euclidean.

Terminologi di dalam $\sum_{i-1}^{n} y_{il} d(X_i, Q_l)$ pada persamaan (1) adalah total biaya *(total cost)* untuk menempatkan *X* ke dalam *clusterl,* seperti halnya dispersi total dari objek di dalam *clusterl* terhadap *prototype* Q_l-nya. E_l diminimalkan jika

$$q_{lj} = \frac{1}{n_l}\sum_{i=1}^{n} y_{il} x_{ij}$$

untuk *j = 1, ..., m* dimana $n_l = \sum_{i=1}^{n} y_{il}$ adalah jumlah objek di dalam *cluster.* Ketika *X* memiliki atribut bertipe *categorikal* , maka diperkenalkan pengukuran kesamaan *(similarity measure)* sebagai berikut :

$$d(X_i, Q_l) = \sum_{j=1}^{m_r} (x_{ij}^r - q_{lj}^r)^2 + \gamma_l \sum_{j=1}^{m_c} \delta(x_{ij}^c, q_{lj}^c) \quad \text{..... (2)}$$

Dimana $\delta(x,q) = 0$ untuk $x = q$ $\delta(x,q) = 1$ untuk $x \neq q$. x_{ij}^r dan q_{lj}^r adalah nilai atribut numerik, sedangkan x_{ij}^c dan q_{lj}^c nilai atribut kategorikal untuk objek ke *i* dan *prototype cluster* ke *l*.m_r dan $m_c$adalah jumlah atribut numerik dan kategorikal. γ_l adalah bobot untuk atribut kategorikal pada *cluster* ke *l.* Sehingga *cost function* E_l bisa dituliskan ulang menjadi sebagai berikut:

$$E_l = \sum_{i=1}^{n} y_{il} \sum_{j=1}^{m_r} (x_{ij}^r - q_{lj}^r)^r + \gamma_l \sum_{i=1}^{n} y_{il} \sum_{j=1}^{m_c} \delta(x_{ij}^c, q_{lj}^c) \quad \text{....(3)}$$

$$E_l = E_l^r + E_l^c \quad \text{.....(4)}$$

Dimana E_l^r adalah biaya total untuk semua atribut numerik dari objek dalam *cluster.*

4. Seleksi

Proses Seleksi, setelah melalui evaluasi fitness maka proses selanjutnya memilih individu yang akan digunakan dalam proses crossover dan selanjutnya akan dimutasikan. Proses seleksi disini menggunakan teknik *Roulette-Wheel.* Pada proses ini menghasilkan dua buah *parent* yang disebut P1 dan P2. *Parent* ini menghasilkan keturunan baru untuk diproses ke tahap *crossover*.

5. Pindah Silang

Operator pindah silang yang digunakan dalam penelitian ini menggunakan satu titik potong, Pada pindah silang satu titik (o*ne-pointcrossover*), satu titik sepanjang kromosom dipilih secara *random*. Segmen induk dari titik point ke kiri atau kekanan ditukar untuk menghasilkan individu baru. sehingga nantinya setiap anaknya akan terdiri dari sebagian pasangan parentnya

6. Mutasi

Tahapan Mutasi ini diperlukan dalam hal mengubah nilai yang ada pada gen didalam kromosom dimana pada proses mutasi tersebut menghasilkan individu baru dengan merubah satu atau lebih gen pada satu individu. Individu yang terpilih untuk proses mutasi dapat dilakukan dengan membandingkan nilai probabilitas mutasinya dengan probabilitas mutasi yang telah ditentukan atau dapat dipilih secara acak *(random)*.

D. Pengujian *Total Cost*

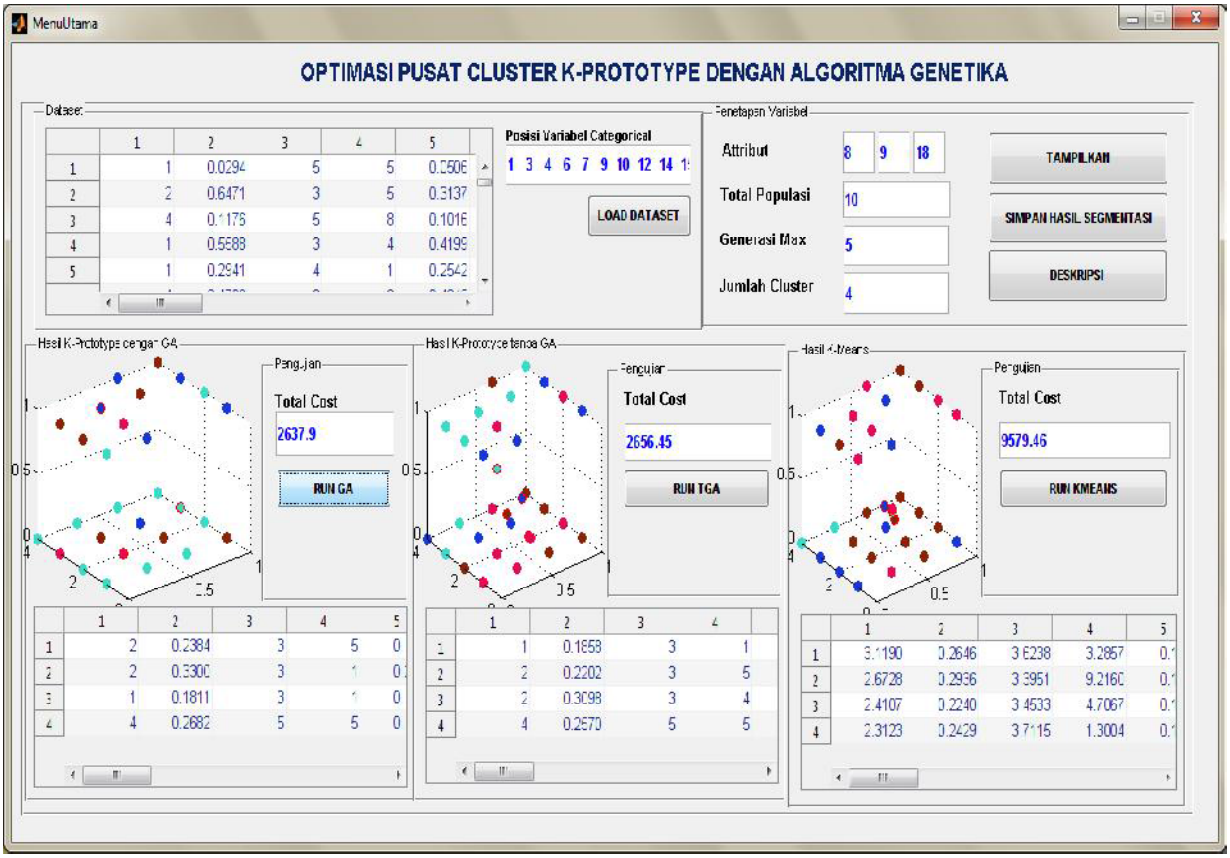
Total *cost* adalah biaya yang dihabiskan untuk menempatkan setiap objekke dalam *cluster* terdekat. Total biaya ini adalah setara dengan total jarak dari setiap objek ke *cluster*yang terbentuk. Semakin kecil biaya yang dihasilkan berarti jarak antara semua obyek dengan *cluster* semakin dekat. Sehingga bisa dikatakan bahwa *cluster* yang terbentuk semakin kompak. *Clustering* sendiri bertujuan tuntuk memperoleh kesamaan ciri-ciri dari setiap objek pada *cluster* tertentu, jika ukuran *cluster* semakin lebar maka kesamaan ciri-ciri yang ada pada sekumpulan objek juga akan semakin rendah, sehingga bisa dikatakan bahwa hasil *cluster* tersebut kurang bagus.

3. HASIL DAN PEMBAHASAN

3.1 Hasil

Untuk mempermudah dalam proses pengelompokan data maka penulis membuat sebuah aplikasi yang dapat digunakan dalam pengujian terhadap metode yang digunakan. Aplikasi pengujian tersebut dapat dilihat pada Gambar 3.

..



Gambar 3. Aplikasi Utama

..

Untuk aplikasi *cluster*ing pada Gambar 3, terdapat beberapa fitur yang diantaranya:

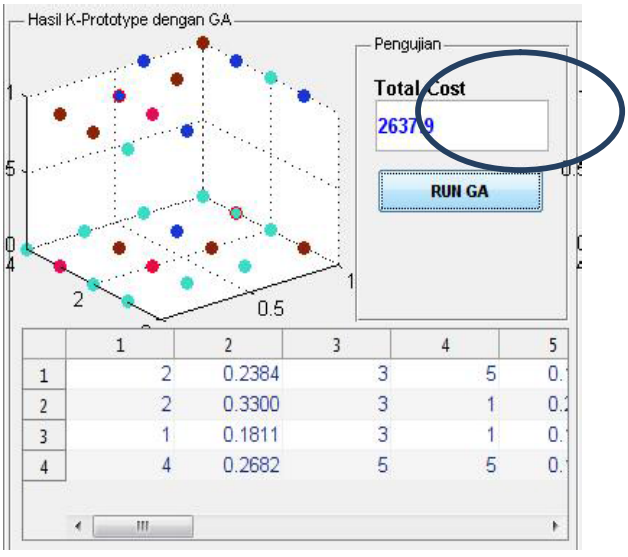
1. Fitur *Dataset*, digunakan untuk memanggil data yang akan diolah, kemudian ditampilkan dalam bentuk Grid, serta menampilkan posisi dari variabel kategorikal.
2. Fitur Penetapan Variabel, saat melakukan input kategori tahap pertama yang dilakukan adalah melakukan penentuan atribut dengan memilih 3 buah atribut bebas dari 20 jumlah atribut yang ada**.** Tahap kedua dilakukan pengisian *textbox* pada Total Populasi yang dapat diisi bebas dengan angka. Total Populasi dapat mempengaruhi hasil dan waktu yang dibutuhkan untuk melakukan proses, apabila total populasi yang digunakan semakin banyak dalam satu generasi, maka akan menghasilkan solusi yang lebih baik. Tahap ketiga, dilakukan pengisian *textbox* pada Generai Max yang dapat diisi bebas dengan angka. Dimana 1 generasi dapat mewakili jumlah populasi yang telah ditentukan pada tahap pengisian total populasi, sehingga

menghasilkan kromosom yang lebih variatif dalam proses fitness. Tahap Keempat melakukan penentuan jumlah *cluster* untuk menggelompokkan data yang diolah sesuai dengan jumlah *cluster* yang diinginkan user.

3. Eksekusi program, Hasil *K-Prototype*dengan Algoritma Genetika. Setelah proses penentuan atribut, maka atribut yang dipilih akan ditampilkan dalam bentuk grafik 3D, menampilkan pusat *cluster* yang telah ditentukan dan ada total jarak antara *cluster* yang terbentuk dilihat dari total *cost* yang digunakan sebagai nilai pengukurannya. Jika nilai total *cost* yang diperoleh semakin kecil berarti jarak antara semua obyek dengan *cluster*nya semakin dekat, Sehingga bisa dikatakan bahwa *cluster* yang terbentuk semakin kompak dan bagus.

3.2 Pembahasan

Untuk mengukur hasil dari metode yang digunakan yaitu Metode *K-Prototype* dengan Algoritma Genetika, maka dilakukan perbandingan dengan metode yang lain. Metode yang digunakan untuk pengujian adalah Metode *Cluster*ing *K-Prototype* tanpa Algoritma Genetika dan Metode *Cluster*ing *K-Means* yang sudah populer. Ketiga metode tersebut sama-sama diuji dengan metode pengukuran hasil *Clustering* disebut dengan *Clustering Criterion*. *Clustering Criterion* inilah yang akan dijadikan nilai *fitness* dari setiap kromosom yang telah dievaluasi. *Clustering Criterion* yang digunakan adalah *cost function* untuk mengukur jarak setiap kelompok data yang terbentuk.Total jarak tiap *cluster* yang nantinya dihasilkan disebut Total*Cost*. Semakin kecil nilai total *cost yang dihasilkan* maka semakin bagus*cluster* yang terbentuk.



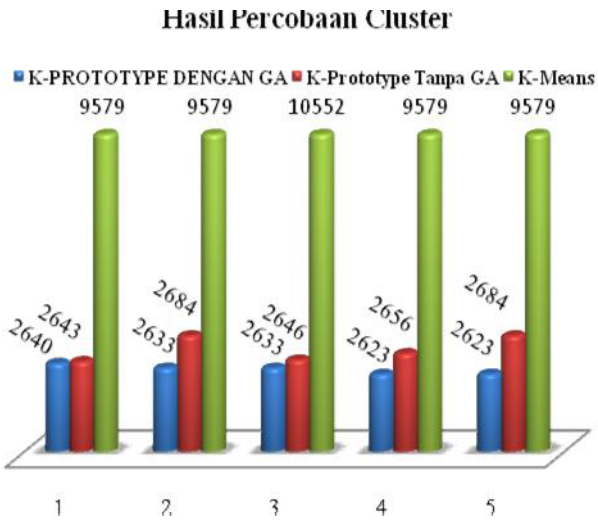
Gambar 4. Hasil K-Prototype dengan GA

Pengujian dilakukan sebanyak 3 kali, dimana pengujian pertama dilakukan dengan menggunakan Inputan Kategori dari Jumlah Pusat *Cluster* sebanyak 4, Generasi Max sebanyak 5 dan Total Populasi sebanyak 30. Hasil pengujian pertama dapat dilihat pada Tabel 1 dan Gambar 5 untuk grafik hasil percobaan pertama.

Tabel 1. Hasil Pengujian 1

Percobaan	Jml Pusat Cluster	*K-Prototype* dengan Ga			Hasil Total *Cost K-Prototype* Tanpa GA	Hasil Total *Cost K-Means*
		Total Populasi	Generasi Max	Hasil Total *Cost*		
1	4	30	5	2640	2643	9579
2	4	30	5	2633	2684	9579
3	4	30	5	2633	2646	10552
4	4	30	5	2623	2656	9579
5	4	30	5	2623	2684	9579

Berdasarkan hasil percobaan pada Tabel 1 dilakukan lima percobaan untuk melihat perbedaan total *cost* untuk tiap-tiap metode dan diperoleh keterangan bahwa *K-Prototype* dengan GA menghasilkan total *cost* terkecil dari pada metode lainnya. Hasil perbandingan yang diperoleh dapat ditabulasikan dalam bentuk diagram batang, seperti yang terlihat pada Gambar 5.



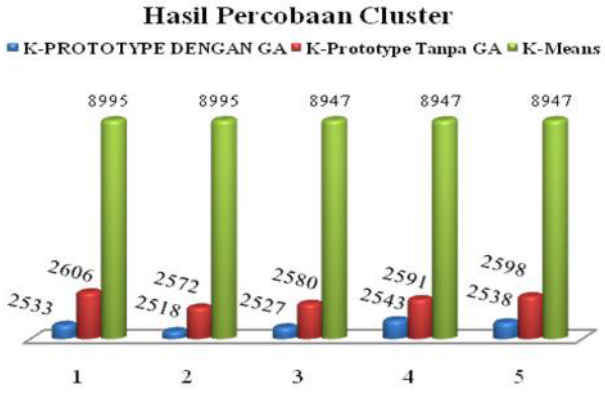
Gambar 5. Grafik Pengujian 1

Hasil pengujian selanjutnya untuk Optimasi *K-Prototype* dengan GA, *K-Protoype* Tanpa GA, dan *K-Means*ditunjukkan seperti padaTabel 2. Sebanyak lima percobaan dilakukan untuk melihat perbedaan total *cost* untuk tiap-tiap metode. dengan menggunakan Inputan Kategori dari Jumlah Pusat *Cluster* sebanyak 5, Generasi Max sebanyak 10 dan Total Populasi sebanyak 30.Berdasarkan hasil yang ditunjukkan pada Tabel 2, diperoleh keterangan bahwa *K-Prototype* dengan GA total *cost* terkecil dari pada metode lainnya. Dan hasil yang diperoleh digambarkan pada grafik pada Gambar 6.

Tabel 2. Hasil Pengujian 2

Percobaan	Jml Pusat Cluster	*K-Prototype* dengan Ga			Hasil Total *Cost K-Prototype* Tanpa GA	Hasil Total *Cost K-Means*
		Total Populasi	Generasi Max	Hasil Total *Cost*		
1	5	30	10	2533	2606	8995
2	5	30	10	2518	2572	8995
3	5	30	10	2527	2580	8947
4	5	30	10	2543	2591	8947
5	5	30	10	2538	2598	8947

Berdasarkan hasil percobaan pada Tabel 2 dengan perubahan pada Jumlah *Cluster* dan Generasi Max yang dilakukan lima percobaan untuk melihat perbedaan total *cost* untuk tiap-tiap metode dan diperoleh keterangan bahwa *K-Prototype* dengan GA menghasilkan total *cost* terkecil dari pada metode lainnya. Hasil perbandingan yang diperoleh dapat ditabulasikan dalam bentuk diagram batang, seperti yang terlihat pada Gambar 6.



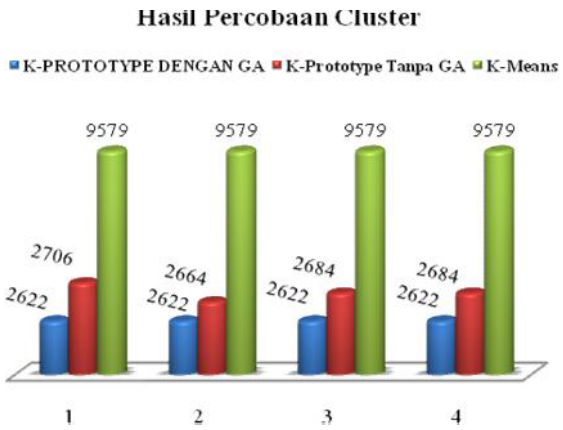
Gambar 6. Grafik Pengujian 2

Berdasarkan data pada Tabel 3, diperoleh keterangan bahwa kenaikan jumlah total populasi dan generasi maksimum akan meningkatkan perbedaan hasil total *cost* untuk *K-Prototype* dengan GA. Menggunakan Jumlah Pusat *Cluster* sebanyak 5, Generasi Max sebanyak 10 dan Total Populasi sebanyak 30.Perbedaan ini terlihat jelas ketika data pada Tabel 2 dan Tabel 3 dibandingkan satu sama lain. Pada Tabel 3, perbandingan total *cost* antara *K-Prototype* dengan GA, *K-Prototype* tanpa GA dan *K-Means* terpaut lebih besar ketika total populasi dan generasi maksimum ditingkatkan.

Tabel 3. Hasil Pengujian 3

Percobaan	Jml Pusat Cluster	*K-Prototype* dengan Ga			Hasil Total *Cost K-Prototype* Tanpa GA	Hasil Total *Cost K-Means*
		Total Populasi	Generasi Max	Hasil Total *Cost*		
1	4	80	40	2622	2706	9579
2	4	80	40	2622	2664	9579
3	4	80	40	2622	2684	9579
4	4	80	40	2622	2684	9579
1	4	80	40	2622	2706	9579

Hasil yang diperoleh pada Tabel 3 dapat ditabulasikan dalam bentuk diagram batang, seperti yang terlihat pada Gambar 8, untuk keperluan perbandingan, visualisasi dalam bentuk diagram batang dapat memperlihatkan perbedaan yang lebih jelas.



Gambar 8. Grafik Pengujian 3

Hasil yang diperoleh pada Tabel 3 dapat ditabulasikan dalam bentuk diagram batang, seperti yang terlihat pada Gambar 8. Pada hasil percobaan ketiga, total *cost* yang dihasilkan dari metode *K-Prototype* dengan Algoritma Genetika lebih sedikit dibandingkan dengan metode uji yang digunakan sebagai pembanding dan pada 4 kali percobaan dengan inputan kategori yang sama, total *cost* yang dihasilkan oleh metode *K-Prototype* dengan GA sama, dapat diartikan bahwa semakin banyak populasi dan generasi max yang digunakan maka dapat memberikan hasil yang lebih maksimal.

Dari proses pengujian yang telah dilakukan sebanyak 3 kali, dapat disimpulkan bahwa metode *K-Prototype* dengan Algoritma Genetika memberikan hasil yang lebih optimal, dilihat dari nilai total *cost* yang dihasilkan lebih kecil dibandingkan dengan *K-Prototype* tanpa Algoritma Genetika dan *K-Means*. Namun untuk mendapatkan hasil yang lebih optimal pada Algortima *K-Prototype* dengan Algoritma Genetika dilakukan peningkatan jumlah populasi dan generasi max sehingga komputasi yang dihasilkan lebih variatif dan hasil yang didapat lebih optimal.

4. KESIMPULAN

Berdasarkan hasil beberapa percobaan yang dilakukan dapat diambil kesimpulan bahwa pada penelitian ini, penggabungan metode *CluteringK-Prototype* dengan Algoritma Genetika dilakukan sebelum fase Evaluasi Fitness. Hal ini berbeda dengan proses Algoritma Genetika pada umumnya, karena setelah proses inisialisasi awal akan langsung dilanjutkan ke fase evaluasi fitness tanpa melalui proses *clustering*. Untuk metode gabungan *K-*

*Prototype*dengan Algortima Genetika yang digunakan dalam penelitian ini menghasilkan optimasi pusat *cluster* dengan hasil yang lebih baik daripada *K-Prototype*tanpa GA, ataupun *K-Means*. Hal ini terlihat dari hasil percobaan yang telah dilakukan, dimana pada saat dilakukan pengujian menggunakan total *cost* atau total jarak metode K-Prototype dengan GA menghasilkan nilai total *cost* terendah yaitu sebesar 2622 kemudian metode K-Prototype tanpa GA menghasilkan nilai total *cost* sebesar 2706 dan Metode K-Means menghasilkan nilai total *cost* sebesar 9579. Oleh karena itu, K-Prototype dengan GA memiliki tingkat kesamaan ciri atau karakteristik dari setiap kelompok yang terbentuk lebih baik.

5. DAFTAR PUSTAKA

[1] Zhexue Huang. (1998), Extensions to the K-Means Algorithm for *Cluster*ing Large Data Sets with Categorical Values, Data Mining and Knowledge Discovery 2, 283–304.

[2] Bashar Al-Shboul, and Sung-Hyon Myaeng. (2009), *Initializing K-Means using Genetic Algorithms*, World Academy of Science, Engineering and Technology 54

[3] Kyoung-jae Kim, Hyunchul Ahn. (2008). *A recommender system using GA K-means clustering in an online shopping market*. Expert Systems with Applications 34. 1200–1209.

[4] Min Feng , Zhenyan Wang. (2011), *A Genetic K-means Clustering Algorithm Based on the Optimized Initial Centers*, Computer and Information Science, Vol. 4, No. 3.

[5] Ming-Yi Shih, Jar-Wen Jheng and Lien-Fu Lai. (2010), *A Two-Step Method for Clustering Mixed Categroical and Numeric Data*, Tamkang Journal of Science and Engineering, Vol.13, No.1, pp.11-19

[6] K.Arun Prabha, R.Saranya. (2011), *Refinement of K-Means Clustering Using Genetic Algorithm*, Journal of Computer Applications (JCA), Volume IV, Issue 2.

[7] Chittu.V,N.Sumathi. (2011), *A Modified Genetic Algorithm Initializing K-Means Clustering*, Global Journal of Computer Science and Technology Volume 11.

[8] Zhexue Huang**.** *Cluster*ing Large Data Sets With Mixed Numeric and Categorical, CSIRO Mathematical and Information Sciences.

[9] Hwei-JenLin, Fu-WenYang and Yang-TaKao. (2005), *An Efficient GA based Clustering Technique,* Tamkang Journal of Science and Engineering, Vol.8, No2, pp.113-122

[10] Duc-Truong Pham, Maria M. Suarez-Alvarez and Yuriy I. Prostov, (2011). *Random search with K-Prototype Algorithm for clustering mixed datasets,* Proceedings Thes Royal Of Society.