BAB 1
PENDAHULUAN

1.1 Latar Belakang

Clustering adalah proses mengelompokkan *record* di dalam *dataset*, dimana kelompok *record* dengan karakteristik yang mirip dikelompokkan ke dalam *cluster* yang sama dan kelompok *record* dengan karakteristik yang berbeda dikelompokkan ke dalam *cluster* yang berbeda(Islam, Estivill-Castro, Rahman, & Bossomaier, 2018). Terdapat beberapa teknik *clustering* yang telah dikemukakan, diantaranya adalah *K-Means*(Lloyd, 1982) yang merupakan salah satu teknik *clustering* yang cukup populer(Wu, et al., 2008).

Pada algoritma *K-Means*, penentuan koordinat titik pusat (*centroid*) mempengaruhi secara langsung kualitas dari proses *clustering*(Maitra, Ghosh, & Peterson, 2010). Penentuan *centroid* yang tidak baik akan berakibat pada diperolehnya hasil *clustering* yang kurang baik(Maulik & Bandyopadhyay, 2000). Penentuan koordinat dari titik pusat (*centroid*) memerlukan sejumlah iterasi, dimana pada awalnya koordinat pusat tiap *cluster* akan ditentukan secara *random* (acak), kemudian tiap *record* pada *dataset* akan ditempatkan ke dalam *cluster* berdasarkan pada kedekatan jarak dengan bilangan acak yang dibangkitkan tersebut(Li, Li, & Wang, 2015).

Penelitian mengenai penentuan *centroid* telah menarik perhatian sejumlah peneliti. (Rahman & Islam, 2012), telah menggunakan pendekatan *fuzzy* untuk menentukan *centroid*, tetapi memiliki keterbatasan pada *dataset* berukuran besar. (Zahra, Ghazanfar, Khalid, Azam, Naeem, & Prugel-Bennett, 2015) telah

mengemukakan metode penentuan *centroid* berbasis *recommender system* yang dilakukan dengan mencari nilai korelasi *pearson* dari sejumlah metode penenuan *centroid* yang telah ada dan mencari nilai yang terdekat dengan permasalahan yang dihadapi. Kelemahan yang ada dari metode ini adalah perlu pendefinisian yang tepat dari *k* jumlah *cluster* dan dapat berakibat pada terjadinya kondisi *local optima*.

Permasalahan *local optima* pada *K-Means* merupakan permasalahan yang umum dihadapi. Oleh karena itu, terdapat permintaan yang besar untuk teknik penentuan *centroid* dan *clustering* yang sederhana tetapi bebas dari keterbatasan *K-Means*. Algoritma Genetika telah digunakan bersama dengan *K-Means* untuk memperbaiki kualitas dari *cluster*(Liu, Wu, & Shen, 2011). (Rahman & Islam, A Hybrid Clustering Technique Combining a Novel Genetic Algorithm with K-Means, 2014) mengemukakan metode *Hybrid Clustering* yang dikenal sebagai GenClust yang menggabungkan pemakaian algoritma *K-Means* dengan Algoritma Genetika. Algoritma Genetika digunakan untuk menentukan jumlah *cluster* dan juga *centroid* dari tiap *cluster*. Namun, yang perlu menjadi pertimbangan adalah percobaan yang dilakukan oleh (Rahman & Islam, A Hybrid Clustering Technique Combining a Novel Genetic Algorithm with K-Means, 2014)menggunakan 50% kromosom yang diperoleh melalui perhitungan *deterministic* dan 50% kromosom diperoleh melalui bilangan acak dan akan mengalami kendala komputasi pada *dataset* berskala besar. (Beg, Islam, & Estivill-Castro, 2016) mengemukakan metode Healthy Population and Multiple Streams Sharing Information (HeMI) yang memodifikasi algoritma GenClust dengan memperbaiki penentuan populasi, dimana metode ini menambah

persentase populasi yang ditentukan melalui perhitungan deterministik dan mengurangi persentase populasi yang ditentukan secara acak.

(Islam, Estivill-Castro, Rahman, & Bossomaier, 2018) telah menyempurnakan metode GenClust dengan menambahkan teknik *modify K-Means (MK-Means*) dan telah menggunakan 100% kromosom yang berkualitas yang diperoleh melalui perhitungan deterministik. Berbeda dengan GenClust, maka pada GenClust++ telah menghindari penggunaan *centroid* yang berasal dari bilangan acak. Penggunakan Algoritma GenClust++ perlu menggarisbawahi beberapa hal yang dapat mempengaruhi kualitas pada algoritma genetika. Dalam algoritma genetika terdapat beberapa parameter penting yang harus didefinisikan yaitu ukuran populasi, proses seleksi, *crossover,* dan mutasi yang harus didefinisikan secara hati-hati agar tidak terjadi konvergensi dini atau lokal optimum yaitu dimana individu-individu dalam populasi konvergen pada suatu solusi optimum lokal sehingga hasil paling optimum tidak dapat ditemukan (Muzid, 2014).

Algoritma GenClust++ yang dikemukakan oleh (Islam, Estivill-Castro, Rahman, & Bossomaier, 2018) telah menunjukkan pengaruh dari proses seleksi terhadap kualitas dari hasil *clustering*. Proses seleksi yang diuji adalah seleksi dengan menggunakan *Roulete Wheel* dan *Tournament Selection*, dan hasil penelitian menunjukkan bahwa proses seleksi menentukan kualitas dari hasil *clustering*.

Perlu dilakukan penelitian lebih lanjut yang mengkaji pengaruh dari beberapa metode *crossover* dan keterkaitannya dengan beberapa metode seleksi

yang lain selain *Roullete Wheel* dan *Tournament Selection* terhadap algoritma *GenClust* sehingga dapat menghasilkan hasil *clustering* yang lebih berkualitas.

Penelitian ini akan membahas mengenai perbandingan antara Algoritma GenClust++ dengan menggunakan beberapa variasi metode *crossover* khususnya di dalam perbandingan untuk mengukur nilai *performance* yang diukur dari *Mean Square Error* yang terjadi untuk hasil *clustering* suatu *dataset*.

1.2 Perumusan Masalah

Algoritma GenClust++ telah dapat meningkatkan kinerja hasil *clustering* dengan algoritma K-Means. Namun, belum ada kajian pengaruh Proses *Crossover* dan Seleksi di dalam algoritma GenClust++. Adapun masalah di dalam penelitian ini adalah Bagaimana Pengaruh dari *Crossover* dan Seleksi terhadap Performa Algoritma GenClust++.

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah dapat meningkatkan kinerja dari Algoritma GenClust++ melalui pemilihan metode *Crossover* dan Seleksi yang sesuai.

1.4 Manfaat Penelitian

Adapun manfaat dari penelitian adalah sebagai berikut.

1. Melalui penelitian ini peneliti akan memperoleh hasil analisis mengenai pengaruh dari metode *crossover* dan seleksi terhadap performance dari algoritma GenClust++.

2. Mengetahui kombinasi proses *crossover* dan seleksi yang terbaik sehingga dapat meningkatkan kualitas hasil *clustering* dengan menggunakan algoritma GenClust++.

1.5 Ruang Lingkup dan Batasan

Sehubungan dengan luasnya permaslahaan dan adanya keterbatasan waktu dan pengetahuan peneliti, maka peneliti membatasi masalah yang akan dibahas di dalam penelitian ini sebagai berikut.

1. *Dataset* yang akan digunakan di dalam penelitian ini adalah *dataset* yang bersumber dari UCI *Machine Learning Repository*.
2. Perhitungan *distance* yang digunakan di dalam algoritma *K-Means* menggunakan *Euclidean Distance*.
3. Perbandingan kinerja di dalam penelitian ini didasarkan pada nilai *Mean Square Error* yang diperoleh pada tiap generasi.

1.6 Sistematika Penulisan

Penulisan tesis ini terdiri dari beberapa bab yang dijabarkan lebih lanjut pada setiap sub bab,sistematika penulisan pada setiap bab sebagai berikut:

BAB I PENDAHULUAN

Pada bab ini yang berisi latar belakang, perumusan masalah, tujuan dan manfaat, ruang lingkup dan sistematika penulisan secara singkat dan jelas.

BAB II LANDASAN TEORI

Pada bab ini menjelaskan teori-teori yang mendasari dan mendukung penulisan tesis ini khsususnya yang berkaitan dengan

algoritma GenClust++ dan teori lain yang mendukung penelitian ini

BAB III METODOLOGI PENELITIAN

Pada bab ini berisi penjelasan lebih mendalam tentang kerangka pikir penelitian, langkah-langkah penelitian, sumber *dataset*, dan metode evaluasi hasil penelitian

BAB IV HASIL DAN PEMBAHASAN

Pada bab ini memuat hasil analisa terhadap pengaruh metode seleksi dan *crossover* terhadap kinerja algoritma *GenClust++*

BABV KESIMPULAN DAN SARAN

Pada bab terakhir ini merupakan kesimpulan dari hasil penelitian yang telah dilakukan dan usulan saran-saran yang berguna untuk penelitian sejenis di masa mendatang.