**JURUSAN TEKNIK INFORMATIKA**

**FAKULTAS TEKNOLOGI INFORMASI**

**INSTITUT TEKNOLOGI SEPULUH NOPEMBER**

**USULAN TUGAS AKHIR**

# **IDENTITAS PENGUSUL**

Nama : **Yoga Bhagawad Gita**

NRP : **5107 100 091**

Dosen Wali : **Wahyu Suadi, S.Kom, MM, M.Kom**

# **JUDUL TUGAS AKHIR**

***“ Implementasi Deteksi Outlier pada Algoritma Hierarchical Clustering “***

# **LATAR BELAKANG**

Pada tugas akhir ini diusulkan sebuah implementasi algoritma *clustering* yang berbasiskan pada algoritma *agglomerative hierarchical clustering*. Pada algoritma *hierarchical clustering* ada dua hal penting yaitu pemilihan jenis pengukuran *similarity* antara obyek dan teknik *linkage.* Pada beberapa kasus *clustering*, penggunaan algoritma *hierarchical clustering* yang berbasiskan *single linkage* menunjukkan hasil yang baik pada berbagai himpunan dataset dengan variasi jumlah dan bentuk. Walaupun begitu, penerapan algoritma ini masih dapat menimbulkan berbagai masalah seperti kepekaan pada adanya outlier dan fluktuasi kepadatan data. Ditambah lagi algoritma ini tidak memungkinkan untuk menentukan jumlah *cluster* secara otomatis. Untuk mengatasi kekurangan-kekurangan itu, maka diperlukan suatu pendekatan baru yang memungkinkan untuk dapat mendeteksi *outlier* dan bahkan dapat secara otomatis menetukan jumlah *cluster* yang dibutuhkan secara tepat.

# **TUJUAN TUGAS AKHIR**

Tujuan dari pembuatan tugas akhir ini adalah meningkatkan performa dari algoritma *hierarchical cluster* yang telah dikembangkan dengan fitur seperti pendeteksi *outlier* dan pendefinisian jumlah *cluster* yang optimal. Dan hasil *clustering* dari algoritma ini akan dibandingkan dengan hasil *clustering* dari algoritma lain untuk mengetahui lebih lanjut mengenai peningkatan performanya.

# **PERMASALAHAN**

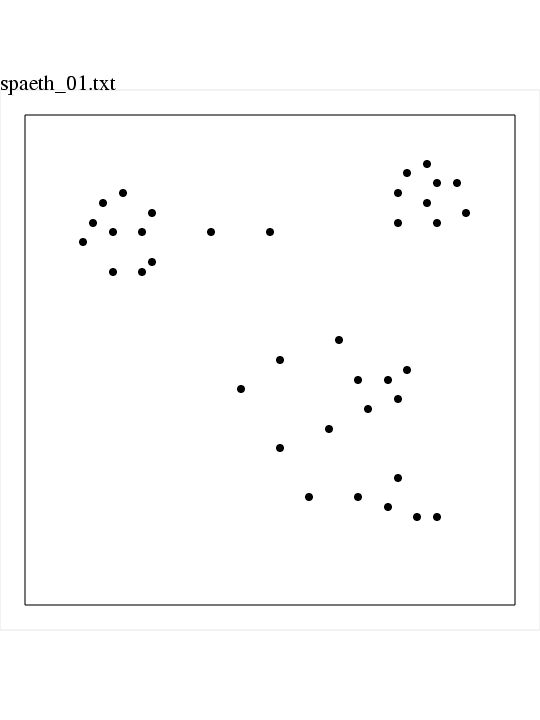
Permasalahan yang diangkat dalam Tugas Akhir ini dapat dijelaskan sebagai berikut:

1. Bagaimana cara menentukan *cluster* tanpa terpengaruh adanya *outlier*.
2. Bagaimana cara untuk mengotomatisasi jumlah *cluster* yang paling tepat pada algoritma *hierarchical cluster*.

# **BATASAN MASALAH**

Permasalahan yang dibahas dalam Tugas Akhir ini memiliki batasan sebagai berikut:

* + - 1. Fitur yang dijadikan acuan adalah yang bertipe numerikal.
      2. Hasil dari tugas akhir ini adalah implementasi dari algoritma *hierarchical clustering* yang telah disempurnakan untuk melakukan *clustering* pada beberapa variasi dataset. Hasil lainnya adalah perbandingan dengan beberapa algoritma *clustering* lainnya yang juga berbasis pada *hierarchical clustering*, yaitu algoritma CURE dan CHAMELEON. Perbandingan dilakukan dengan menggunakan dataset yang terdapat pada situs <http://people.sc.fsu.edu/~jburkardt/datasets/spaeth/spaeth.html> dan <http://people.sc.fsu.edu/~jburkardt/datasets/spaeth2/spaeth2.html>.

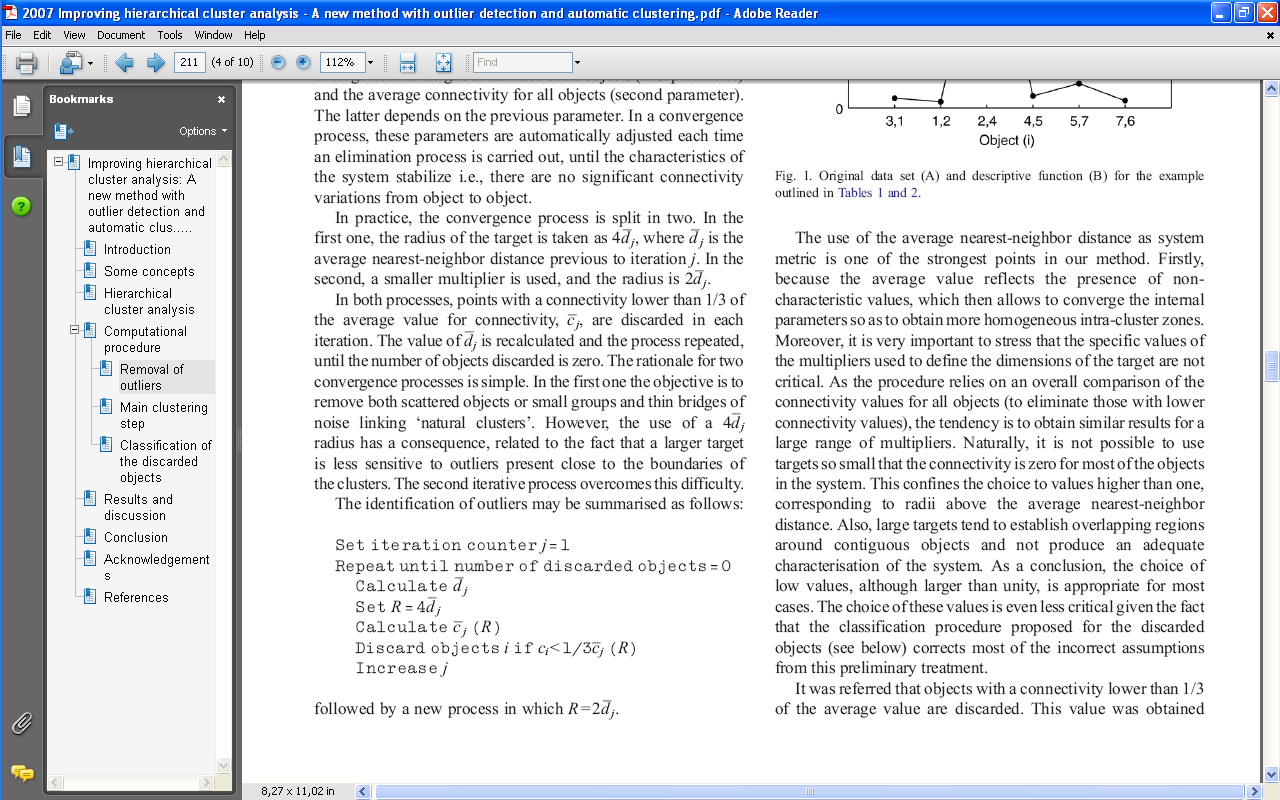


Gambar 1. Contoh dataset yang akan diujikan pada algoritma *clustering*.

# **RINGKASAN TUGAS AKHIR**

Pada tugas akhir ini diimplementasikan algoritma pengembangan dari algoritma *hierarchical clustering* dengan tiga tahap, yaitu tahap penghilangan *outlier*, pengelompokan *cluster*, dan pengklasifikasian *outlier*.

Tahap penghilangan outlier merupakan tahap awal yang harus dilakukan. Tahap ini dilakukan agar proses pengelompokan *cluster* hanya terfokus dengan data asli tanpa terpengaruh oleh kehadiran data *outlier*. Berikut adalah *pseudocode* untuk menghilangkan outlier pada suatu dataset.



dj adalah nilai jarak terdekat dua buah titik dari dua *cluster* yang akan bergabung pada saat dilakukan proses *agglomerative hierarchical clustering*, dan cj adalah nilai *connectivity*-nya.

Tahap berikutnya adalah tahap pengelompokan *cluster* dengan pendeteksian jumlah *cluster* yang optimal, maksudnya adalah sebuah proses clustering akan otomatis menyesuaikan jumlah *cluster* yang dibutuhkan dengan berdasar pada pola data inputannya. Tahap ini memanfaatkan nilai-nilai hasil *linkage* untuk mendapatkan variabel baru yang disebut *descriptive function* (DF).

DFi,i+1 = d2i,i+1 (1)

Nilai d adalah jarak terpendek antara dua titik yang merupakan hasil perhitungan di tahap *linkage*. Selanjutnya dengan memanfaatkan nilai DF dari setiap pasang titik, didapatkan nilai *threshold* yang cocok untuk menentukan jumlah cluster yang tepat untuk sebuah dataset.

DFsep > 6 x (Q3 – Q­1) (2)

Dengan Q1 dan Q3 masing-masing adalah nilai kuartil pertama dan ketiga dari nilai pada *descriptive function*.

Tahap yang terakhir dilakukan adalah pengklasifikasian *outlier* yang telah dihilangkan pada tahap pertama. Tahap ini tidak harus dilakukan. Setelah *cluster* terbentuk maka data yang dianggap sebagai *outlier* pada tahap pemrosesan yang pertama dimasukkan kembali dengan mengklasifikasikannya pada beberapa kelompok *cluster* yang telah terbentuk pada tahap pemrosesan kedua. Metode yang digunakan pada tahap ini adalah metode *supervised clustering* yaitu dengan menggunakan *k-nearest neighbour*.

Secara garis besar *flowchart* dari algoritma ini adalah sebagai berikut:



Gambar 2. *Flowchart* algoritma modifikasi *hierarchical clustering*

# **METODOLOGI**

Ada beberapa tahap dalam proses pengerjaan tugas akhir ini. Berikut tahap-tahap dalam pembuatannya :

1. **Studi literatur**

Pada tahap ini akan dipelajari sejumlah literatur mengenai konsep dan teknologi yang akan digunakan. Literatur yang digunakan meliputi paper referensi, buku referensi, ebook dan dokumentasi internet, diantaranya tentang:

* *Outlier Detection*
* *Hierarchical Clustering Analysis*
* *Supervised Clustering*

1. **Pengimplemetasian rumus secara analitis**

Tahap ini merupakan tahap penerapan dasar teori yang telah dipahami dalam penggunaan rumus dan algoritma yang akan diterapkan.

1. **Perancangan perangkat lunak dan desain sistem**

Pada tahap ini dilakukan proses perancangan perangkat lunak, berdasarkan literatur yang telah dikaji kemudian di buat desain model data, diagram alir proses-proses yang ada, dan desain antar muka aplikasi.

1. **Pengimplementasian perangkat lunak**

Pada tahap ini dilakukan proses pengimplemetasian, dengan berdasar pada rancangan awal perangkat lunak dan literature yang telah dikaji. Aplikasi telah mulai dibuat secara menyeluruh.

1. **Uji coba dan evaluasi**

Pada tahap ini akan diuji aplikasi yang telah selesai diimplemetasikan tersebut. Pengujian dan evaluasi akan dilakukan pada sejumlah dataset yang memiliki variasi dimensi dan kepadatan data.

1. **Penyusunan laporan tugas akhir**

Pada tahap ini disusun buku sebagai dokumentasi dari pelaksanaan tugas akhir, yang mencakup seluruh konsep, teori, implementasi, serta hasil yang telah dikerjakan. Laporan tugas akhir ini akan dibagi menjadi beberapa bab sebagai berikut:

1. Bab I, Pendahuluan, berisi latar belakang, permasalahan, tujuan, batasan permasalahan, metodologi, dan sistematika penulisan.
2. Bab II, Landasan Teori, akan dibahas dasar ilmu yang mendukung pembahasan tugas akhir ini.
3. Bab III, Desain Aplikasi.
4. Bab IV, Implementasi dari aplikasi yang telah dibuat, akan dilakukan pembuatan aplikasi yang dibangun dengan komponen-komponen yang telah ada yang sesuai dengan permasalahan dan batasannya yang telah dijabarkan pada bab pertama.
5. Bab V, Uji coba dan analisa hasil, akan dilakukan uji coba berdasarkan parameter-parameter yang ditetapkan, dan kemudian dilakukan analisa terhadap hasil uji coba tersebut.
6. Bab VI, Penutup, berisi kesimpulan yang dapat diambil dari Tugas Akhir ini beserta saran untuk pengembangan selanjutnya
7. **DAFTAR PUSTAKA**

[1] J. A. S. Almeida, L. M. S. Barbosa, A. A. C. C. Pais, S. J. Formosinho, *Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering*, Departemento de Quimica, Universidade de Coimbra, 3004-535 Coimbra, Portugal, Chemometrics and Intelligent Laboratory Systems 87 (2007) 208–217.

[2] George Karypis, Eui-Hong Han, Vipin Kumar, *Chameleon: A Hierarchical Clustering Algorithm Using Dynamic Modelling*, Departement of Computer Science and Engineering, University of Minnesota, 4-192 EECS Bldg., 200 Union St. SE Minneapolis, MN 55455, USA, IEEE Comput. 32 (8) (1999) 68–75.

[3] S. Guha, R. Rsatogi, K. Shim, *Cure: an efficient clustering algorithm forlarge databases*, Stanford University, Stanford, CA 94305, Bell Laboratories, Murray Hill, NJ 07974, Inf. Syst. 26 (1) (2001) 35–58.

1. **JADWAL PEMBUATAN TUGAS AKHIR**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Kegiatan** | **Bulan** | | | | | | | |
| **1** | | **2** | | **3** | | **4** | |
| Studi Pustaka |  |  |  |  |  |  |  |  |
| Perancangan Perangkat Lunak |  |  |  |  |  |  |  |  |
| Pembuatan Perangkat Lunak |  |  |  |  |  |  |  |  |
| Uji Coba dan Evaluasi |  |  |  |  |  |  |  |  |
| Analisa Hasil Uji Coba |  |  |  |  |  |  |  |  |
| Penyusunan Buku |  |  |  |  |  |  |  |  |

# **LEMBAR PENGESAHAN**

**Surabaya, 4 Oktober 2010**

Menyetujui,

Dosen Pembimbing I

**( Ahmad Saikhu, S.Si, M.Kom )**

**( NIP. 197107182006041001)**