Pendekatan Algoritma Divide and Conquer pada Hierarchical Clustering

Agnes Theresia Damanik / 13510100[1]
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
[1]13510100@std.itb.ac.id

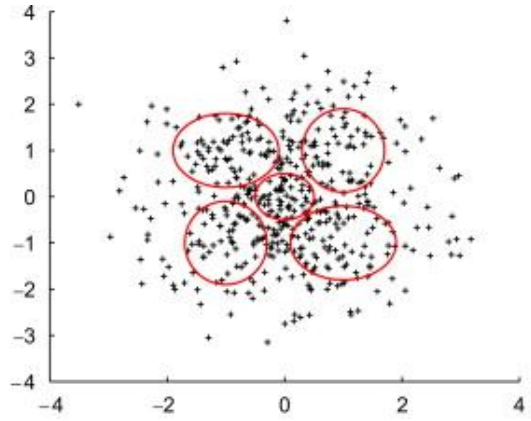
***Abstract*—Clustering adalah pengelompokan data berdasarkan kemiripan data. Clustering memudahkan individu untuk mencari data dan menemukan kesamaan yang terdapat antar data. Hierarchical clustering adalah salah satu pendekatan yang digunakan untuk melakukan clustering. Divide and conquer merupakan algoritma yang dapat mengatasi persoalan yang memiliki karakteristik yang mirip dengan upa-persoalan tersebut.**

***Index Terms*—Clustering, Divide and Conquer, mirip, Hierarchical Clustering.**

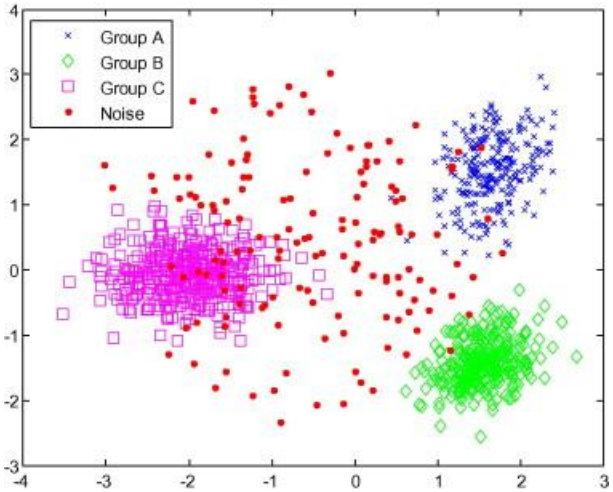
I. PENDAHULUAN

Clustering adalah metode panganalisaan data yang bertujuan mengelompokkan data dengan karakteristik yang sama ke dalam suatu 'wilayah' yang sama dan data dengan karakteristik yang berbeda ke dalam 'wilayah' yang lain. Beberapa contoh clustering yang telah dilakukan beberapa di antaranya adalah novel video face clustering, data stream clustering pada rekaman telepon, multimedia data, serta transaksi finansial.

Terdapat beberapa pendekatan untuk mengembangkan metode clustering. Dua pendekatan utama adalah clustering dengan pendekatan partisi (*partition-based clustering*) dan clustering dengan pendekatan hirarki (*hierarchical clustering*). Clustering dengan pendekatan partisi dilakukan dengan mengelompokkan data dengan memilah-milah data yang dianalisas ke dalam cluster-cluster yang ada. Clustering dengan pendekatan hirarki merupakan clustering yang melakukan pengelompokan data dengan membuat suatu hirarki berupa dendogram, data yang mirip akan ditempatkan pada hirarki yang berdekatan dan yang tidak ditempatkan pada hirarki yang berjauhan.



Gambar 1 Data Dua Dimensi dengan 5 Cluster



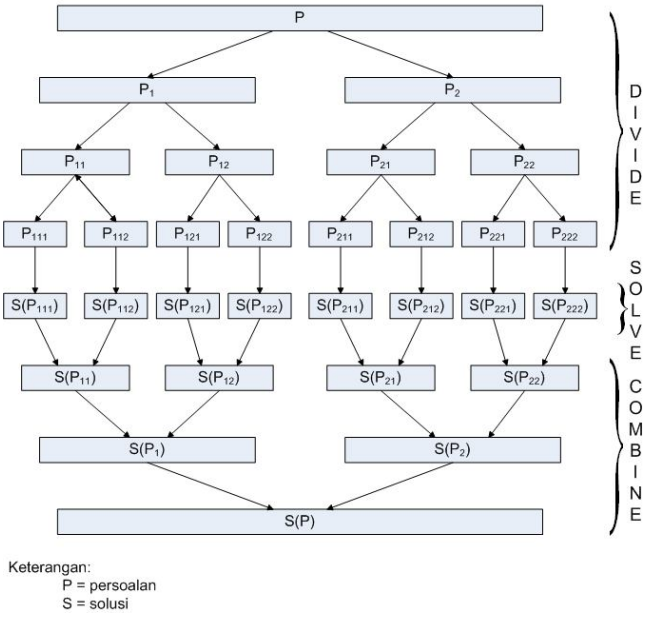
Gambar 2 Obyek Dua Dimensi untuk Clustering Ecoli data

II. DASAR TEORI

2.1 DIVIDE AND CONQUER

Karakteristik utama dari sebuah permasalahan algoritma divide and conquer adalah persoalan tersebut, apabila dibagi ke dalam beberapa upa-masalah, memiliki karakteristik yang sama dengan karakteristik masalah

asal sehingga persoalan seringkali dapat diatasi dengan skema rekursif. Pada algoritma divide and conquer persoalan yang ada dibagi menjadi beberapa upa-masalah yang memiliki kemiripan dengan persoalan semula namun berukuran lebih kecil lalu upa-masalah tersebut masing-masing diselesaikan (dipecahkan) secara rekursif setelah semua dpecahkan solusi dari setiap upa-masalah digabungkan sehingga membentuk solusi persoalan semula.



Gambar 3 Ilustrasi Algoritma Divide and Conquer

Objek persoalan yang dibagi merupakan instans persoalan yang berukuran n seperti table (larik), matriks, serta eksponen. Metode divide and conquer secara natural dapat diungkapkan sebagai skema rekursif. Skema umum dari suatu algoritma divide and conquer diperlihatkan oleh gambar di berikut ini.

```
procedure DIVIDE_and_CONQUER(input n : integer)
{
Menyelesaikan masalah dengan algoritma D-and-C
Masukan: masukan yang berukuran n
Keluaran: solusi dari masalah semula
}
Deklarasi
  r, k : integer
Algoritma
  if n ≤ n0 then
{ukuran masalah sudah cukup kecil }
     SOLVE upa-masalah yang berukuran n ini
  else
     Bagi menjadi r upa-masalah, masing-masing berukuran n/k
     for masing-masing dari r upa-masalah do
        DIVIDE_and_CONQUER(n/k)
     endfor
     COMBINE solusi dari r upa-masalah menjadi solusi masalah semula }
  endif
```

Gambar 4 Skema Umum Algoritma Divide and Conquer

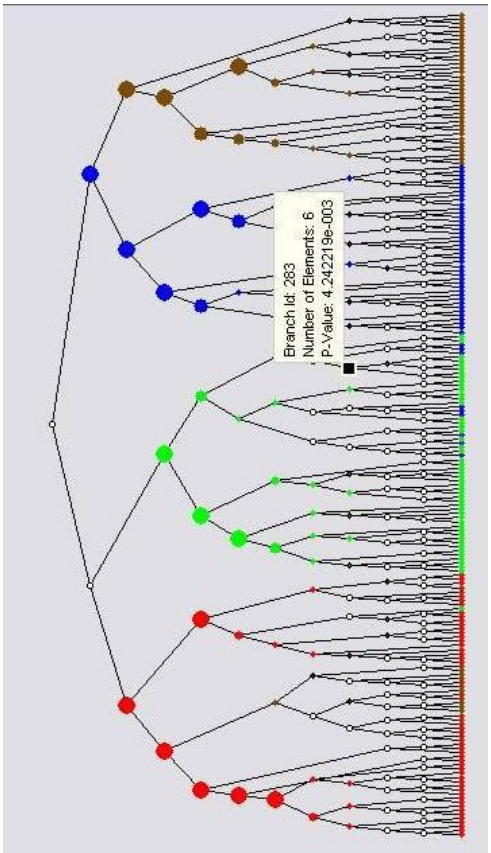
2.2 HIERARCHICAL CLUSTERING

Seperti yang telah disebutkan sebelumnya, hierarchical clustering membangun sebuah hirarki berupa dendogram untuk mengelompokkan data-data yang mirip dan menjauhkan data-data yang tidak mirip. Strategi yang digunakan dalam hierarchical clustering terdiri atas dua jenis yaitu sebagai berikut.

- Agglomerative
  Pendekatan '*bottom up*' dimana pengamatan diawali pada masing-masing cluster dan setiap pasangan cluster dilebur menjadi satu menaik pada bagian atas hirarki
- Divisive
  Pendekatan '*top down*' dimana pengamatan dimulai dari suatu cluster dan pemecahan dilaksanakan secara rekursif seiring dengan satu pecahan bergerak ke bawah hirarki

Pada umumnya, penggabungan dan pemecahan yang disebutkan di atas diputuskan secara *greedy*. Hasil dari hierarchical clustering direpresentasikan dalam dendogram. Dendogram adalah diagram pohon yang biasanya digunakan untuk mengilustrasikan pengaturan dari cluster-cluster yang dihasilkan oleh pendekatan hierarchical clustering.

Untuk kasus-kasus umum, kompleksitas dari strategi agglomerative adalah $O(n^3)$ yang tidak efektif digunakan pada himpunan data yang berukuran besar.



Gambar 5 Hasil Hierarchical Clustering

III. PENERAPAN ALGORITMA

Algoritma divide and conquer dapat diterapkan untuk melakukan clustering dan menghasilkan hasil yang hirarki dari cluster tersebut namun bukan hierarchical clustering.

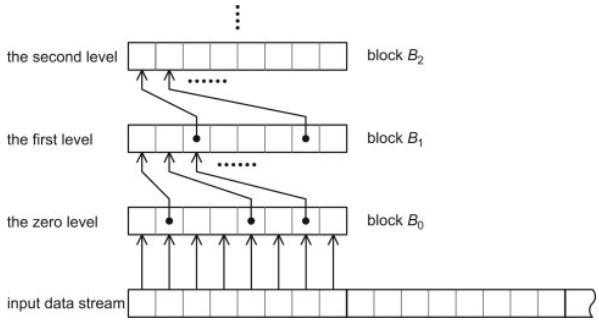
Pendekatan divide and conquer yang dilakukan adalah dengan cara sebagai berikut.

1. Bagi (divide) himpunan data ke dalam bagian-bagian yang memiliki ukuran sama besar
2. Cluster (mengelompokkan) masing-masing bagian tersebut secara tersendiri akhirnya
3. Lebur bagian-bagian yang sudah di-*cluster* tadi menurut ketentuan jarak sebagai berikut:

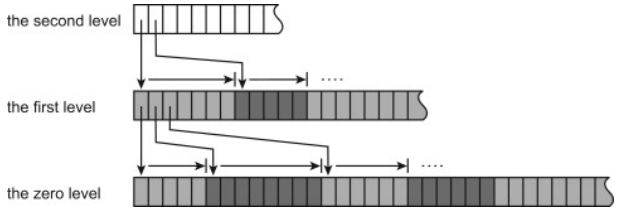
   jika $d(C1,C2)<t$; maka: lebur $C1$ dan $C2$

Berikut ini merupakan proses cluster yang terjadi di harddisk.

Divide and conquer clustering. Data stream diisikan pada blok B_0 sepotong demi sepotong.



Susunan storage dari hirarki cluster pada harddisk. Cluster yang berada pada level yang lebih tinggi memiliki pointer ke keseluruhan cluster pada level di bawahnya.



Algoritma dalam melakukan clustering dengan menggunakan pendekatan hierarchical clustering diperlihatkan pada gambar 6.

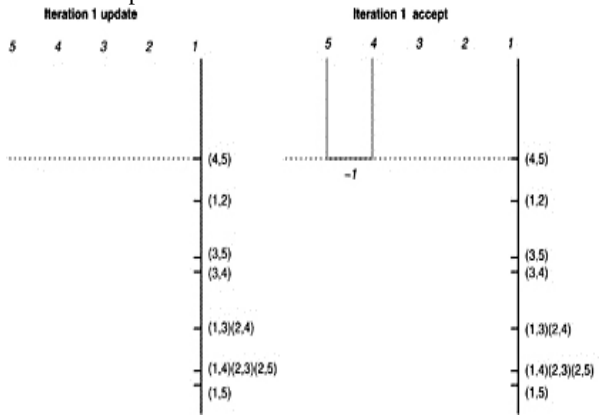
Algorithm 1 The algorithm of direct k-means clustering
Input: N objects in the R-dimensional space, and M (number of sub-cluster means)
Output: Initial M sub-clusters that contain all N objects
initialize M first means $r_0^{(1)}, r_0^{(2)}, \ldots, r_0^{(M)}$
repeat
  for each input object r_i, $1 \leqslant i \leqslant N$
    do
      assign r_i to D_j with nearest mean $r_0^{(j)}$, such as $\varphi(r_i, r_0^{(j)}) \leqslant \varphi(r_i, r_0^{(u)}), 1 \leqslant j, u \leqslant M$
  for each sub-cluster D_j, $1 \leqslant j \leqslant M$
  do
    recalculate the mean (centroid) of objects $r_i \in D_j$, $r_0^{(j)} = \sum_{i \in D_j} r_i / |D_j|$, where $|D_j|$ defines the cardinality of D_j
  compute the quality function G
until no object has changed sub-clusters (or G does not change)

Gambar 6 Algoritma Hierarchical Clustering

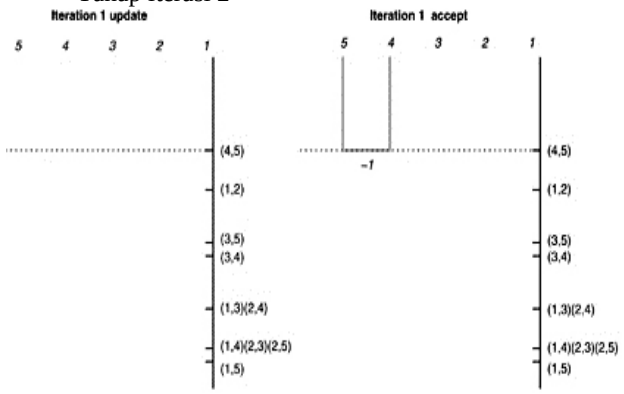
Tahapan pendekatan hierarchical clustering diperlihatkan pada gambar-gambar di bawah.

Pada tahap inisialisasi (iterasi t=0), dibuat lima terminal node yang diberi label 1...5 secara berurutan. Pada setiap tahap iterasi dimulai dari t=1 dan memenuhi t≥1 dilakukan juga tahap untuk me-*update* yang akan diikuti oleh tahap penerimaan dimana dibuat node internal yang baru. Pada waktu update, dilakukan update atas sejumlah node yang sekiranya mampu memenuhi persyaratan.

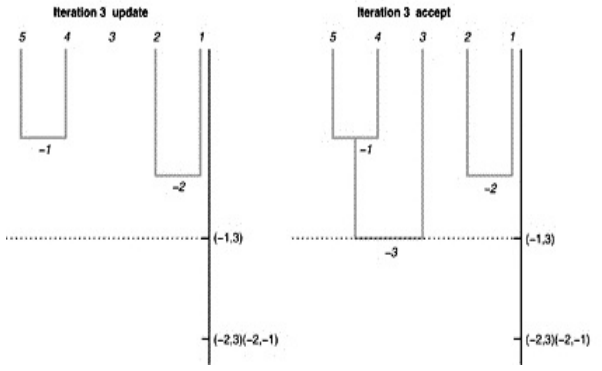
Tahap iterasi 1



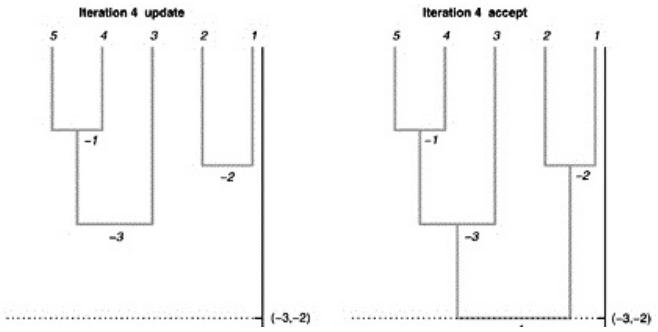
Tahap iterasi 2



Tahap iterasi 3



Tahap iterasi 4



IV. HASIL PENGUJIAN

Pada dasarnya, solusi clustering yang diperoleh dengan menggunakan algoritma divide and conquer dan pendekatan hirarki cukup mirip. Salah satu hal yang membedakan adalah kompleksitas waktu yang disebabkan.

Manurut algoritma divide and conquer kompleksitas waktu ketika menyelesaikan masalah clustering adalah $\boldsymbol{O(p)+O(pq)}$ dengan p sebagai jumlah bagian-bagian (potongan-potongan) dan q merupakan rata-rata dari jumlah cluster pada setiap bagian-bagian.

Pada algoritma single link hierarchical clustering kompleksitas waktu yang diperoleh adalah $\boldsymbol{O(n^2)}$. Namun jika ada asumsi bahwa ketika hierarchical clustering diterapkan dan masing-masing bagian membutuhkan waktu yang sama dan konstan untuk semua bagian-bagian maka $O(n^2)$ akan menjadi $O(1)$.

V. KESIMPULAN

Salah satu cara menyelesaikan permasalahan clustering adalah melalui pendekatan hirarki (hierarchical clustering). Pendekatan algoritma divide and conquer juga dapat digunakan untuk menyelesaikan permasalahan clustering dan dalam beberapa kasus himpunan data berukuran besar, algoritma ini cukup mangkus jika dibandingkan pendekatan hierarchical clustering.

REFERENCES

[1] Munir, Rinaldi. 2009. Diktat Kuliah Strategi Algoritma IF3051

[2] http://yudiagusta.wordpress.com/clustering/ diakses Jumat 21 Desember 2011 pukul 13.00 WIB

[3] http://www.sciencedirect.com/science/article/pii/S0925231207000719 diakses Jumat 21 Desember 2012 pukul 14.30 WIB

[4] http://en.wikipedia.org/wiki/Hierarchical_clustering diakses Jumat 21 Desember 2012 pukul 14.41 WIB

[5] http://cstheory.stackexchange.com/questions/10769/divide-and-conquer-approach-for-hierarchical-clustering diakses Jumat 21 Desember 2012 pukul 15.01 WIB

[6] http://www.nature.com/hdy/journal/v103/n1/fig_tab/hdy200929f1.html diakses Jumat 21 Desember 2012 pukul 15.33 WIB

PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 21 Desember 2012

ttd

Agnes Theresia Damanik
13510100