BAB II

TINJAUAN PUSTAKA

2.1 Pengertian *Data Mining*

Data mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam *database*. *Data mining* adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar. (Turban, dkk. 2005)

Definisi umum dari *data mining* itu sendiri adalah proses pencarian pola-pola yang tersembunyi (*hidden patern*) berupa pengetahuan (*knowledge*) yang tidak diketahui sebelumnya dari suatu sekumpulan data yang mana data tersebut dapat berada di dalam *database*, data *werehouse*, atau media penyimpanan informasi yang lain. Hal penting yang terkait di dalam *data mining* adalah:

1. *Data mining* merupakan suatu proses otomatis terhadap data yang sudah ada.
2. Data yang akan diproses berupa data yang sangat besar.
3. Tujuan *data mining* adalah mendapatkan hubungan atau pola yang mungkin memberikan indikasi yang bermanfaat.

   (Kusrini dan Taufiq, 2009)

Data mining dilakukan dengan *tool* khusus, yang mengeksekusi operasi *data mining* yang telah didefinisikan berdasarkan model analisis. *Data mining* merupakan proses analisis terhadap data dengan penekanan menemukan informasi yang tersembunyi pada sejumlah data besar yang disimpan ketika menjalankan bisnis perusahaan. Kemajuan luar biasa yang terus berlanjut dalam bidang *data mining* didorong oleh beberapa faktor antara lain:

1. Pertumbuhan yang cepat dalam kumpulan data.
2. Penyimpanan data dalam data *warehouse*, sehingga seluruh perusahaan memiliki akses ke dalam *database* yang andal.
3. Adanya peningkataan akses data melalui navigasi web dan internet.
4. Tekanan kompetisi bisnis untuk meningkatkan penguasaan pasar dalam globalisasi ekonomi.
5. Perkembangan teknologi perangkat lunak untuk *data mining* (ketersediaan teknologi.
6. Perkembangan yang hebat dalam kemampuan komputasi dan pengembangan kapasitas media penyimpanan.

   (Larose, 2005)

Istilah *data mining* dan *knowledge discovery in databases* (KDD) sering kali digunakan secara bergantian untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu basis data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan satu sama lainnya. Salah satu tahapan dalam keseluruhan proses KDD adalah *data mining*. Proses KDD itu ada 5 tahapan yang dilakukan secara terurut, yaitu:

1. Data *selection*

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang digunakan untuk proses *data mining*, disimpan dalam suatu berkas, terpisah dari basis data operasional.

2. Pre-*processing* / *cleaning*

Sebelum proses *data mining* dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data yang menjadi fokus KDD. Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data.

3. Transformation

Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*. Proses *coding* dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

4. *Data mining*

Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam *data mining* sangat bervariasi. Pemilihan metode atau

algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keselurahan.

5. *Interpretation* / *evalution*

Pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagaian dari proses KDD yang disebut interpretation. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang d*item*ukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya. (Fayyad, 1996)

2.2 Pengelompokan *Data Mining*

Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu:

1. Deskripsi

Terkadang peneliti dan analis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecendrungan yang terdapat dalam data. Sebagai contoh, petugas pengumpul suara mungkin tidak menemukan keterangan atau fakta bahwa siapa yang tidak cukup profesional akan sedikit didukung dalam pemilihan presiden. Deskripsi dari pola dan kecendrungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecenderungan.

2. Estimasi

Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik dari pada ke arah kategori. Model dibangun dengan *record* lengkap menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi.

3. Prediksi

Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang. Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi dapat pula digunakan (untuk keadaan yang tepat) untuk prediksi.

4. Klasifikasi

Dalam klasifikasi, terdapat target variabel kategori. Sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah.

5. Pengklusteran

Pengklusteran merupakan pengelompokan *record*, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. Kluster adalah kumpulan *record* yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan dengan *record-record* dalam kluster lain.

Pengklusteran berbeda dengan klasifikasi yaitu tidak adanya variabel target dalam pengklusteran. Pengklusteran tidak mencoba untuk melakukan klasifikasi, mengestimasi, atau memprediksi nilai dari variabel target. Akan tetapi, algoritma pengklusteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan *record* dalam satu kelompok akan bernilai maksimal, sedangkan kemiripan dengan *record* dalam kelompok lain akan bernilai minimal.

6. Asosiasi

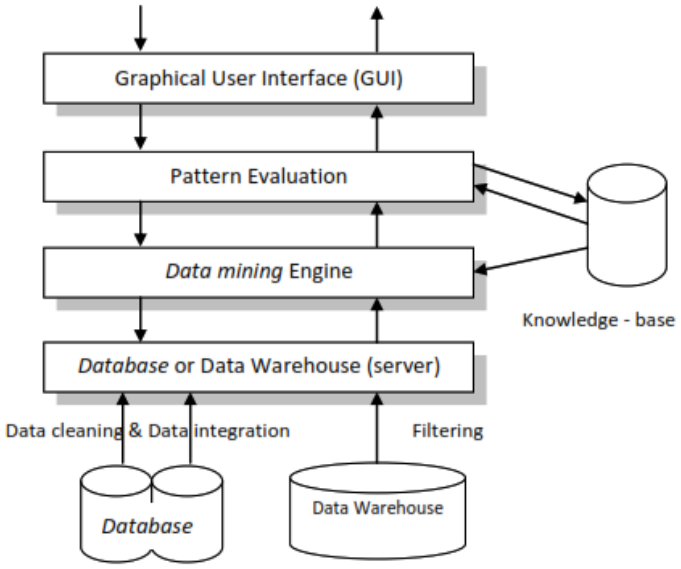
Tugas asosiasi dalam *data mining* adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja (*market basket analysis*). (Larose,2005)

2.3 Arsitektur *Data Mining*

Arsitektur utama dari sistem *data mining*, pada umumnya terdiri dari beberapa komponen sebagai berikut:

1. *Database*, data *warehouse*, atau media penyimpanan informasi, terdiri dari satu atau beberapa *database*, data *warehouse*, atau data dalam bentuk lain. Pembersihan data dan integrasi data dilakukan terhadap data tersebut.
2. *Database*, data *warehose*, bertanggung jawab terhadap pencarian data yang relevan sesuai dengan yang diinginkan pengguna atau *user*.
3. Basis pengetahuan (*Knowledge Base*), merupakan basis pengetahuan yang digunakan sebagai panduan dalam pencarian pola.

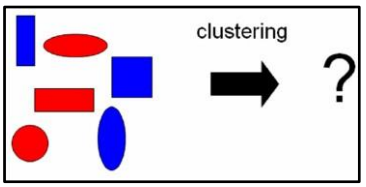
4. *Data mining engine*, merupakan bagaian penting dari sistem dan idealnya terdiri dari kumpulan modul-modul fungsi yang digunakan dalam proses karakteristik (*characterization*), klasifikasi (*clasiffication*), dan analisis kluster (*cluster analysis*). Dan merupakan bagian dari *software* yang menjalankan program berdasarkan algoritma yang ada.
5. Evaluasi pola (*pattern evaluation*), komponen ini pada umumnya berinteraksi dengan modul-modul *data mining*. Dan bagian dari *software* yang berfungsi untuk menemukan *pattern* atau pola-pola yang terdapat dalam *database* yang diolah sehingga nantinya proses *data mining* dapat menemukan *knowledge* yang sesuai.
6. Antar muka (*Graphical user interface*), merupakan modul komunikasi antara pengguna atau user dengan sistem yang memungkinkan pengguna berinteraksi dengan sistem untuk menentukan proses *data mining* itu sendiri.



Gambar 2.1 Arsitektur *Data mining*

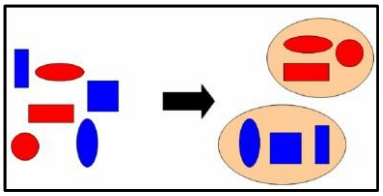
2.4 *Cluster*

Cluster adalah sekumpulan objek yang mempunyai "Kesamaan" diantara anggotanya dan memiliki "Ketidaksamaan" dengan objek lain pada *cluster* lainnya, dengan kata lain sebuah *cluster* adalah sekumpulan objek yang digabung bersama karena persamaan atau kedekatannya. *Clustering* adalah proses membuat pengelompokkan sehingga semua anggota dari setiap partisi mempunyai persamaan berdasarkan matrik tertentu. Gambar 2.2 berikut menunjukkan contoh data yang akan dilakukan klasterisasi.



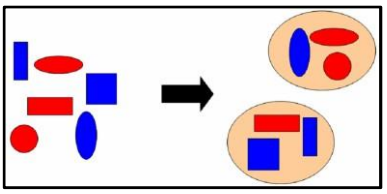
Gambar 2.2 Data Sebelum di Klasterisasi

Jika data dilakukan *clustering* (pengelompokkan) berdasarkan warna, maka pengelompokkannya seperti yang terlihat pada gambar 2.3.



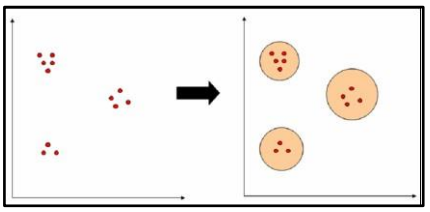
Gambar 2.3 Klasterisasi Berdasarkan Similaritas (Kesamaan) Warna

Jika data dilakukan *clustering* (pengelompokkan) berdasarkan bentuk, maka pengelompokannya dapat dilihat seperti gambar 2.4:



Gambar 2.4 Klasterisasi Berdasarkan Similaritas (Kesamaan) Bentuk

Selain dengan menggunakan similaritas (kesamaan) berdasarkan bentuk dan warna, *clustering* juga bisa dilakukan dengan menggunakan similaritas berdasarkan jarak, artinya data yang memiliki jarak berdekatan akan membentuk satu *cluster*, contohnya seperti dapat dilihat pada gambar 2.5:



Gambar 2.5 Klasterisasi Berdasarkan Similaritas (Kesamaan) Jarak

Ada beberapa perbedaan antara metode klasifikasi dan metode *clustering*, dimana pada dasarnya terdapat tiga poin perbedaan yaitu : data, label dan analisa hasil. Perbedaan tersebut dapat ditabelkan seperti tabel 2.1 berikut:

Tabel 2.1 Perbedaan Klasifikasi dan Klasterisasi

Perbedaan	Klasifikasi	Klasterisasi
Data	*Supervised*	*Unsupervised*
Label	Ya	Tidak
Analisa Hasil	*Error Ratio*	*Variance*

Data *supervised* pada klasifikasi artinya data melalui pembelajaran terbimbing, sedangkan data *unsupervised* pada klasterisasi artinya data tidak melalui pembelajaran terbimbing. Analisa hasil pada klasterisasi dinyatakan dengan *variance* yang menunjukkan variansi data dalam satu *cluster*, sedangkan klasifikasi analisa hasil diukur menggunakan rasio kesalahan (*error ratio*). Pada dataset yang digunakan oleh klasifikasi terdapat satu *attribut* (label) yang berfungsi sebagai *attribut* target, sedangkan dataset pada klasterisasi tidak terdapat attribut (label) sebagai attribut target.

2.4.1 Karakteristik *Clustering*

Ada beberapa karakteristik dari *cluster*ing, masing-masing akan dijelaskan berikut ini:

1. *Partitioning Clustering*
    a. Disebut juga *exclusive clustering*
    b. Setiap data harus termasuk dalam *cluster* tertentu
    c. Memungkinkan bagi setiap data yang termasuk *cluster* tertentu pada suatu tahapan proses, pada tahapan berikutnya berpindah ke *cluster* yang lain.

Contoh : *K-Means*, residual analysis.

2. *Hierarchical Clustering*

   a. Setiap data harus masuk ke dalam *cluster* tertentu

   b. Suatu data yang masuk kedalam *cluster* tertentu pada suatu tahapan proses, tidak dapat berpindah ke *cluster* lain.

Contoh: *Single Linkage, Centroid Linkage, Average Linkage dan Complete Linkage*

2.5 Metode Pengelompokan

Metode pengelompokan pada dasarnya ada dua, yaitu metode pengelompokan Hirarki (*Hirarchical Clustering Method*) dan metode non Hirarki (*Non Hirarchical Clustering Method*). Metode pengelompokan hirarki digunakan apabila belum ada informasi jumlah kelompok yang akan dipilih. Sedangkan metode pengelompokan Non Hirarki bertujuan untuk mengelompokkan *n* objek kedalam *k* kelompok (k<n), dimana nilai *k* telah ditentukan sebelumnya. Salah satu prosedur pengelompokan pada Non Hirarki adalah dengan menggunakan metode *k-means*. Metode ini merupakan metode pengelompokan yang bertujuan untuk mengelompokkan objek sedemikian hingga jarak tiap-tiap objek ke pusat kelompok didalam suatu kelompok adalah minimum.

2.6 *Distance Space*

Distance Space berfungsi untuk menghitung jarak antara data dan *centroid*. Ada beberapa macam *distance space* yang sudah diimplementasikan salah satunya adalah : *Euclidean distance space*.

Euclidean sering digunakan karena penghitungan jarak dalam *distance space* ini merupakan jarak terpendek yang bisa didapatkan antara dua titik yang diperhitungkan.

Jarak antara dua titik dapat dihitung dengan cara:

$$d = |x - y| = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

Dimana:

d : jarak

p : dimensi data

x : titik data pertama,

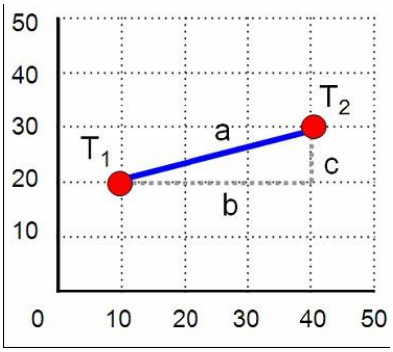
y : titik data kedua,

Dalam *Euclidean* perhitungan yang dilakukan merupakan jarak terpendek antara dua titik. Jika ada *n* titik pengamatan dengan *p* variabel, maka sebelum dilakukan pengelompokkan data atau objek, terlebih dahulu menentukan ukuran kedekatan sifat antar data.

Ukuran data yang bisa digunakan adalah euclidius (*euclidian distance*), antara dua titik dari *p* dimensi pengamatan. Jika antar titik X (x1,x2,...,xn) dan titik Y=(y1,y2,...yn) di tentukan dengan rumus:

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 ..... + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

Penghitungan jarak dengan *Euclidian Distance* untuk dua titik seperti diilustrasikan pada gambar 2.6:



$$a^2 = b^2 + c^2$$
$$a = \sqrt{b^2 + c^2}$$
$$a = \sqrt{(40-10)^2 + (30-20)^2}$$
$$a = \sqrt{(30)^2 + (10)^2}$$
$$a = \sqrt{900 + 100}$$
$$a = \sqrt{1000} = 31.628$$

Gambar 2.6 Penghitungan Jarak Dua Titik

Dari penggambaran diatas dapat diartikan bahwa semakin kecil jarak atau nilai *d*, maka semakin besar keserupaan antar objek tersebut.

2.7 Analisa *Cluster*

Analisis *cluster* adalah suatu analisis statistik yang bertujuan memisahkan obyek kedalam beberapa kelompok yang mempunyai sifat berbeda antar kelompok yang satu dengan yang lain. Dalam analisis ini tiap-tiap kelompok bersifat homogen antar anggota dalam kelompok atau variasi obyek dalam kelompok yang terbentuk sekecil mungkin.

2.7.1 Proses Analisis *Cluster*

Tujuan utama analisis *cluster* menggabungkan objek-objek yang mempunyai kesamaan ke dalam sebuah kelompok atau *cluster*. Untuk mencapai tujuan itu kita harus menjawab tiga pertanyaan, yaitu :

1. Bagaimana kita mengukur tingkat kesamaan ?

   Ada tiga ukuran untuk mengukur kesamaan antar objek, yaitu ukuran korelasi, ukuran jarak, dan ukuran asosiasi.

2. Bagaimana kita membentuk *cluster* ?

   Prosedur yang diterapkan harus dapat mengelompokkan objek-objek yang memiliki kesamaan yang tinggi ke dalam suatu *cluster* yang sama.

3. Berapa banyak *cluster* yang akan kita bentuk ?

   Pada prinsipnya jika jumlah *cluster* berkurang maka homogenitas alam *cluster* secara otomatis akan menurun.

2.7.2 Proses Pengambilan Keputusan

Pengambilan keputusan dengan analisis *cluster* memiliki 6 tahapan, yaitu : menentukan tujuan analisis *cluster*, menentukan desain penelitian analisis *cluster*, menentukan asumsi analisis *cluster*, menurunkan *cluster-cluster* dan memperkirakan *overall fit*, menginterpretasi hasil analisis *cluster*, mengukur tingkat validasi hasil analisis *cluster*.

Langkah 1 : Tujuan Analisis *Cluster*

Tujuan analisis *cluster* secara khusus antara lain :

1. Pengelompokkan Analisis *cluster* digunakan dengan tujuan *explanatory* maupun *confirmatory*
2. Penyederhanaan data Analisis *cluster* menetapkan struktur dari observasi atau data bukan variabel.
3. Pengidentifikasian hubungan Analisis *cluster* dapat menunjukkan ada tidaknya hubungan antar observasi atau obyek dalam analisis.

Langkah 2 : Desain Penilitian dalam Analisis *Cluster*

A. Mendeteksi *Outliers*

Dalam melakukan pemilahan obyek kedalam *cluster-cluster*, analisis tidak hanya peka terhadap variabel-veriabel yang tidak sesuai dengan kasus yang diteliti tetapi juga peka terhadap *outliers* (obyek-obyek yang "berbeda" dengan obyek yang lainnya). *Outliers* terjadi karena 2 dua hal, yaitu :

1. Observasi "menyimpang" yang tidak mewakili populasi
2. Suatu *undersampling* kelompok-kelompok dalam populasi yang menyebabkan *underrepresentation* kelompok-kelompok dalam sampel.

Dalam kedua kasus tersebut, *outliers* dapat mengubah struktur sebenarnya dari populasi sehingga kita akan memperoleh *cluster-cluster* yang tidak sesuai dengan struktur sebenarnya dari populasi tersebut. Karena itu, pembuangan *outliers* sangat penting dalam analisis ini. *Outliers* dapat dilihat melalui

Profile Diagram. Outliers adalah obyek-obyek dengan profil-profil yang berbeda, atau *value* yang berbeda dalam satu atau beberapa variabel.

B. Kesamaan Ukuran

Konsep kesamaan yang diperlukan dalam analisis *cluster*. *Interobject Similarity* adalah sebuah ukuran untuk kesesuaian atau kemiripan, diantara objek-objek yang akan dipilah menjadi beberapa *cluster*. *Interobject Similarity* dapat diukur dengan beberapa cara, antara lain : *Correlatioal Measures, Distance Measures,* dan *Association Measures*. Pemilihan metode tergantung pada tujuan dan jenis data. *Correlatioal Measures* dan *Distance Measures* digunakan untuk data dengan tipe *metic*, sedangkan *Association Measures* digunakan bila data bertipe *non-metic*.

C. Standarisasi Data

Sama halnya dengan seleksi kesamaan ukuran, dalam standarisasi data ni peneliti harus menjawab sebuah pertanyaan, yaitu : Apakah data yang tersedia harus distandarisasi? Dalam menjawab pertanyaan ini, penelti harus memperhatikan beberapa masalah, misalnya, jarak nilai dari masing-masing variabel karena perbedaan skala. Secara umum, variabel dengan penyebaran nilai yang tinggi mempunyai dampak yang lebih pada hasil akhir. Karena itu, peneliti diharapkan mengetahui secara lengkap pengukuran dari variabel-variabel. Proses standarisasi dalam analisi *cluster* ada dua, yaitu : standarisasi berdasarkan variabel dan standarisasi berdasarkan observasi.

Langkah 3 : Asumsi-asumsi Analisis *Cluster*

Dalam analisis *cluster*, peneliti harus lebih memperhatikan masalah : seberapa besar sampel mewakili populasi (*representativeness*) dan ada tidaknya *multicollinearity*.

Langkah 4 : Menurunkan *Cluster-Cluster* dan Memperkirakan *Overall Fit*

Peneliti pertama kali harus menentukan *clustering algorithm* yang akan digunakan untuk membentuk *cluster* dan selanjutnya memutuskan berapa *cluster* yang akan dibentuk. Dua hal ini mempunyai implikasi yang substensial tidak hanya pada hasil yang akan diperoleh tetapi juga pada intepretasi hasil tersebut.

Langkah 5 : Interpretasi *Cluster*

Tahap interpretasi meliputi pengujian masing-masing *cluster* dalam terminology macam *cluster* untuk menamai atau memberikan keterangan secara tepat sebagai gambaran sifat dari *cluster*. Ketika memulai proses interpretasi, ada satu ukuran yang sering digunakan yaitu *cluster centroid.*

Jika prosedur pengelompokan dilakukan terhadap data asli, maka ini akan memberikan gambaran yang logic. Tetapi jika data telah distandarisasi atau jika analisis *cluster* dilakukan dengan menggunakan hasil analisis faktor (faktor komponen), peneliti harus mengembalikan skor asli untuk variabel asal dan menghitung rata-rata profiles menggunakan data ini.

Gambaran dan interpretasi *cluster*, memberikan hasil lebih daripada deskriptif. *Pertama,* Metode ini memberikan sebuah rata-rata untuk perkiraan masing-masing *cluster* yang terbentuk sebagaimana yang dikemukakan pada toeri sebelumnya atau pengalaman praktek. *Kedua,* Gambaran *cluster* memberikan jalan untuk membuat perkiraan signifikansi praktis. Peneliti mungkin memerlukan bahwa perbedaan substansi yang ada pada sejumlah variable *cluster* dan penyelesaian *cluster* akan dikembangkan sampai tampak sejumlah perbedaan.

Langkah 6 : Validasi dan Gambaran *Cluster*

Analisis *cluster* agak bersifat subjektif dalam penentuan penyelesaian *cluster* yang optimal, sehingga peneliti seharusnya memberikan perhatian yang besar mengenai validasi dan jaminan tingkat signifikansi pada penyelesaian akhir dari *cluster*. Meskipun tidak ada metode untuk menjamin validitas dan tingkat signifikansi, beberapa pendekatan telah dikemukakan untuk memberikan dasar bagi perkiraan peneliti.

A. Validasi Hasil *Cluster*

Validasi termasuk usaha yang dilakukan oleh peneliti untuk menjamin bahwa hasil *cluster* adalah representatif terhadap populasi secara umum, dan dengan demikian dapat digeneralisasi untuk objek yang lain dan stabil untuk waktu tertentu. Pendekatan langsung dalam hal ini adalah dengan analisis sample secara terpisah kemudian membandingkan antara hasil *cluster* dengan perkiraan masing-masing *cluster*. Pendekatan ini sering tidak praktis, karena

adanya keterbatasan waktu dan biaya atau ketidaktersediaan objek untuk perkalian analisis *cluster*. Dalam hal ini pendekatan yang biasa digunakan adalah dengan membagi sample menjadi dua kelompok. Masing- masing dianalisis *cluster* secara terpisah, kemudian hasinya dibandingkan.

B. Profiling Hasil *Cluster*

Tahap *Profiling* meliputi penggambaran karakteristik masing-masing *cluster* untuk menjelaskan bagaimana mereka bisa berbeda secara relevan pada tiap dimensi. Tipe ini meliputi penggunaan analisis diskriminan. Prosedur dimulai setelah *cluster* ditentukan. Peneliti menggunakan data yang sebelumnya tidak masuk dalam prosedur *cluster* untuk menggambarkan karakteristik masing-masing *cluster*. Meskipun secara teori tidak masuk akal (rasional) dalam perbedaan silang *cluster*, akan tetapi hal ini diperlukan untuk memprediksi validasi taksiran, sehingga minimal penting secara praktek.

2.8 *K-Means*

Algoritma *K-Means* adalah Metode *clustering* non *hierarchical* berbasis jarak yang membagi data kedalam *cluster* dan algoritma ini bekerja pada attribut numerik. Algoritma *K-Means* termasuk dalam partitioning *clustering* yang memisahkan data ke *k* daerah bagian yang terpisah. Algoritma *K-Means* sangat terkenal karena kemudahan dan kemampuannya untuk mengklaster data besar dan *outlier* dengan sangat cepat.

K-Means merupakan metode klasterisasi yang paling terkenal dan banyak digunakan di berbagai bidang karena sederhana, mudah diimplementasikan, memiliki kemampuan untuk mengklaster data yang besar, mampu menangani data outlier, dan kompleksitas waktunya linear O(nKT) dengan n adalah jumlah dokumen, K adalah jumlah kluster, dan T adalah jumlah iterasi. Dalam algoritma *K-Means*, setiap data harus termasuk ke *cluster* tertentu pada suatu tahapan proses, pada tahapan proses berikutnya dapat berpindah ke *cluster* yang lain. Pada dasarnya penggunaan algoritma *K-Means* dalam melakukan proses *clustering* tergantung dari data yang ada dan konklusi yang ingin dicapai. Untuk itu digunakan algoritma *K-Means* yang didalamnya memuat aturan sebagai berikut:

A. Jumlah *cluster* yang perlu di inputkan
B. Hanya memiliki attribut bertipe numerik

Algoritma *K-Means* pada awalnya mengambil sebagian dari banyaknya komponen dari populasi untuk dijadikan pusat *cluster* awal. Pada step ini pusat *cluster* dipilih secara acak dari sekumpulan populasi data. Berikutnya *K-Means* menguji masing-masing komponen didalam populasi data dan menandai komponen tersebut ke salah satu pusat *cluster* yang telah di definisikan tergantung dari jarak minimum antar komponen dengan tiap-tiap pusat *cluster*. Posisi pusat *cluster* akan dihitung kembali sampai semua komponen data digolongkan kedalam tiap-tiap *cluster* dan terakhir akan terbentuk posisi *cluster* baru.

Algoritma *K-Means* pada dasarnya melakukan 2 proses yakni proses pendeteksian lokasi pusat *cluster* dan proses pencarian anggota dari tiap-tiap *cluster*.

Proses algoritma *K-Means* :

1. Tentukan *k* sebagai jumlah *cluster* yang ingin dibentuk
2. Bangkitkan *k* centroids (titik pusat *cluster*) awal secara random.
3. Hitung jarak setiap data ke masing-masing *centroids*.
4. Setiap data memilih *centroids* yang terdekat
5. Tentukan posisi *centroids* baru dengan cara menghitung nilai rata-rata dari data-data yang terletak pada *centroids* yang sama.
6. Kembali ke langkah 3 jika posisi centroids baru dengan centroids lama tidak sama.

Berdasarkan cara kerjanya, algoritma *K-Means* memiliki karakteristik:

1. *K-Means* sangat cepat dalam proses *clustering*
2. *K-Means* sangat sensitif pada pembangkitan *centroids* awal secara random
3. Memungkinkan suatu *cluster* tidak mempunyai anggota
4. Hasil *clustering* dengan *K-Means* bersifat tidak unik (Selalu berubah- ubah) – terkadang baik, terkadang jelek.

Adapun tujuan dari data *cluster*ing ini adalah untuk meminimalisasikan *objective function* yang diset dalam proses *cluster*ing, yang pada umumnya berusaha meminimalisasikan variasi di dalam suatu *cluster* dan memaksimalisasikan

variasi antar *cluster*. Ada dua cara pengalokasian data kembali ke dalam masing-masing *cluster* pada saat proses iterasi *cluster*ing.

K-Means dalam pengalokasian data ke dalam masing-masing *cluster* dapat dilakukan dengan 2 cara yaitu Hard *K-Means* dan Fuzzy K-Means. Perbedaan dari kedua metode tersebut terletak pada asumsi yang dipakai sebagai dasar dari pengalokasian data. Hard disini dalam artian suatu data secara tegas atau pasti dinyatakan sebagai anggota satu *cluster* tertentu dan tidak menjadi anggota *cluster* yang lain. Sedangkan Fuzzy diartikan masing-masing data mempunyai nilai kemungkinan untuk dapat bergabung ke setiap *cluster* yang ada.

2.9 Average Linkage Clustering

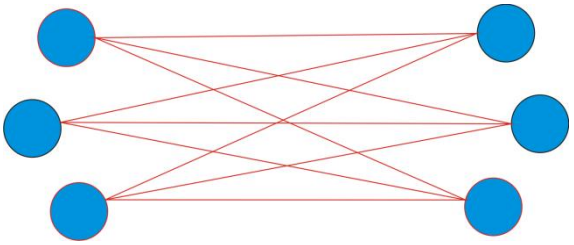
Average Linkage adalah proses pengelompokan yang didasarkan pada jarak rata-rata antar objeknya. Prosedur ini hampir sama dengan single linkage maupun complete linkage, namun kriteria yang digunakan adalah rata-rata jarak seluruh individu dalam suatu cluster dengan jarak seluruh individu dalam cluster yang lain. (Kusrini 2007).

Average Linkage memperlakukan jarak antara dua cluster sebagai jarak rata-rata antara semua pasangan item-item dimana satu anggota dari pasangan tersebut kepunyaan tiap cluster. Mulai dengan mencari matriks jarak D = {dik} untuk memperoleh objek-objek paling dekat (paling mirip) misalnya U dan V. Objek-

objek ini digabungkan untuk membentuk cluster (UV). Untuk langkah dari algoritma diatas jarak-jarak antara (UV) dan cluster W yang lain di tentukan oleh:

$$d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_w}$$

dimana d_{ik} adalah jarak antara objek *i* dalam *cluster (UV)* dan objek *k* dalam *cluster W* dan N_{uv} dan N_w berturut-turut adalah banyaknya item-item dasar *cluster (UV)* dan *W*.



Gambar 2.7 Ilustrasi Average Linkage Clustering