BAB II

TINJAUAN PUSTAKA

2.1 Analisis Cluster

Menurut Tan, 2006 analisis cluster adalah sebuah proses untuk mengelompokan data ke dalam beberapa cluster atau kelompok sehingga data dalam satu cluster memiliki tingkat kemiripan yang maksimum dan data antar cluster memiliki kemiripan yang minimum. Sedangkan menurut Mattjik dan Sumertajaya (2011), Analisis cluster merupakan teknik peubah ganda yang mempunyai tujuan utama untuk tmengelompokkan objek-objek berdasarkan kemiripan karakteristik yang dimilikinya Penggerombolan objek tersebut dilakukan berdasarkan peubah yang diamati pada n objek. Ukuran kemiripan diukur dengan menggunakan ukuran jarak. Salah satu ukuran jarak yang dapat digunakan adalah jarak Euclid (Mattjik dan Sumertajaya 2011). Hasil dari anlisis cluster dapat di pengaruhi oleh objek yang di kelompokan, ukuran kemiripan/ketidak miripan, skala ukuran, dan metode clustering yang digunakan.

1) Ukuran kemiripan

Kemiripan antar pasangan objek x dan y di notasikan sebagai $si\ (x,y)$.Dalam menentukan mirip tidaknya suatu objek, dapat di lihat dari besarnya nilai $si\ (x,y)$. Pada objek yang memiliki kemiripan maka nilai $si\ (x,y)$ akan besar dan sebaliknya apabila nilai $si\ (x,y)$ kecil makan objek merupakan pasangan yang tidak mirip.

Unruk setiap pasangan objek x dan y berlaku 3 kondisi berikut (Kandardzic, 2011):

1) $0 \leq si \ \ (x, y) \leq 1$, kemiripan bernilai 0 dan 1.
2) $si\ (x, y) = 1$, setiap objek mirip dengan dirinya sendiri.
3) $si\ \ (x, y) = si\ (y, x)$, kemiripan besifat simetri.

2) Ukuran ketidakmiripan

Ukuran ketidakmiripan digunakan untuk mencari jarak antara pasangan objek di dalam data. Jarak antara pasangan objek x dan y dinyataka n dengan $d(x, y), d(x, y)$ akan bernilai besar jika x dan y merupakan pasangan objek yang tidak mirip, sebaliknya $d(x, y)$ akan bernilai kecil jika x dan y merupakanpasangan objek yang mirip. Untuk setiap objek x dan y berlaku kondisi berikut (Han & Kamber, 2001):

1) $d(x, y) \geq 0$, jarak merupakan bilangan non- negatif.
2) $d(x, y) = 0$, jarak suatu objek dengaan dirinya sendiri = 0.
3) $d(x, y) = d(y, x)$, jarak bersifat simetri.

Semakin besar nilai ukuran ketidakmiripan antara dua objek maka semakin besar pula perbedaan antara kedua objek tersebut, sehingga makin cenderung untuk tidak berada dalam kelompok yang sama (Johnson & Wichern, 2007).

2.2 Cluster Ensemble

Strehl dan Gosh (2002) memperkenalkan sebuah metode yang digunakan untuk mengombinasikan sekumpulan solusi gerombol yang disebut *Cluster*

Ensemble. Metode *Cluster Ensemble* memiliki keunggulan dibanding metode penggerombolan lain. Penelitian yang dilakukan oleh Strehl dan Gosh (2002) menunjukkan bahwa metode *Cluster Ensemble* mampu meningkatkan kualitas dan kekekaran solusi gerombol. Tantangan untuk mendapatkan solusi gerombol dengan kualitas yang baik dan adanya keragaman solusi gerombol yang dihasilkan dari metode yang berbeda merupakan motivasi dikembangkannya *Cluster Ensemble*.

Penggerombolan pada *Cluster Ensemble* dilakukan dengan mengombinasikan berbagai solusi dari berbagai metode penggerombolan hingga diperoleh satu penggerombolan akhir yang lebih baik. Input yang dibutuhkan adalah solusi penggerombolan yang telah diperoleh dengan menggunakan berbagai hasil penggerombolan tanpa melihat karakteristik data awal. Secara umum, penggerombolan objek dengan metode *cluster ensemble* dilakukan dalam dua tahap menurut Iam-on dan Garret (2010), yaitu:

1. Membentuk anggota *ensemble* yang anggotanya adalah solusi dari berbagai metode penggerombolan yang berbeda.
2. Mengombinasikan seluruh anggota *ensemble* untuk memperoleh satu solusi akhir yang dinamakan fungsi *Consensus*.

2.3 Pengelompokan Data Numerik

Ada dua jenis data clustering yang sering dipergunakan dalam proses pengelompokan data yaitu hierarchical (hierarki) data clustering dan non hierarchical (non hierarki) data clustering. Penelitian ini menggunakan metode non hirarki sebagai metode pengelompokan data numerik. Metode ini dimulai dengan

proses penentuan jumlah cluster terlebih dahulu. Terdapat beberapa metode non hirarki yang dapat d gunakan, salah satunya *k-mean* dan *k-medoid.*

- Algoritma K-Means

Algoritma K-Means merupakan algoritma yang relatif sederhana untuk mengklasifikasikan atau mengelompokkan sejumlah besar obyek dengan atribut terentu ke dalam kelompok-kelompok sebanyak K. Metode ini menjadi salah satu metode data clustering non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih kelompok. Algoritma K-Means pertama kali diperkenalkan oleh MacQueen JB pada tahun 1976. Pada algoritma K-Means jumlah cluster K telah ditentukan terlebih dahulu.

Kelebihan Algoritma K-Means diantaranya adalah mampu mengelompokkan objek besar dan pencilan obyek dengan sangat cepat sehingga mempercepat proses pengelompokan. Adapun kekurangan yang dimiliki oleh K-Means sangat sensitif pada pembangkitan titik pusat awal secara random, hasil pengelompokan bersifat tidak unik (selalu berubah-ubah), Algoritma K-Means walaupun proses pengerjaannya cepat tetapi keakuratannya tidak dijamin.

Cara kerja algoritma K-Means:

1. Tentukan K sebagai jumlah cluster yang ingin dibentuk,
2. Bangkitkan K centroid (titik pusat cluster) awal secara random
3. Dalam menentukan buah pusat cluster awal dilakukan pembangkitan bilangan random yang merepresentasikan urutan data input. Pusat awal

cluster didapatkan dari data sendiri bukan dengan menentukan titik baru, yaitu dengan merandom pusat awal dari data.

4. Hitung jarak setiap data ke masing-masing centroids, untuk mengukur jarak antara data dengan pusat cluster digunakan Euclidian Distance.

- Algoritma K-Medoid

Dalam metode K-medoid setiap cluster dipresentasikan dari sebuah objek di dalam cluster yang disebut dengan medoid. Tujuannya adalah menemukan kelompok K-cluster (jumlah cluster) diantara semua objek data di dalam sebuah kelompok data. Clusternya dibangun dari hasil mencocokkan setiap objek data yang paling dekat dengan cluster yang dianggap sebagai medoid sementara. Langkah-langkah menghitung medoids, yaitu:

1. Pilih point k sebagai inisial centroid/nilai tengah (medoids) sebanyak k cluster.
2. Cari semua point yang paling dekat dengan medoid, dengan cara menghitung jarak vector antar dokumen. (menggunakan Euclidian distance)
3. Secara random, pilih point yang bukan medoid.
4. Hitung total distance
5. If TD baru < TD awal, tukar posisi medoid dengan medoids baru, jadilah medoid yang baru.
6. Ulangi langkah 2 – 5 sampai medoid tidak berubah.

2.4 Pengelompokan Data Kategorik

Metode *clustering* yang digunakan untuk tipe data kategorik adalah algoritma *ROCK*. *ROCK* pertama kali diperkenalkan oleh Guha, Rastogi, & Shim pada tahun 1999. Metode *ROCK* menggunakan konsep *link* sebagai ukuran kemiripan untuk membentuk *cluster*-nya. Metode *ROCK* dapat menangani *outlier* dengan cukup efektif. Pemangkasan *outlier* memungkinkan untuk membuang yang tidak ada tetangga, sehingga titik tersebut tidak berpartisipasi dalam pengelompokan. Namun dalam beberapa situasi, *outlier* dapat hadir sebagai *cluster-cluster* yang kecil (Guha, Rastogi, & Shim, 1999).

Clustering untuk data kategorik dengan algoritma *ROCK* dilakukan dengan tiga langkah. Adapun langkahnya yaitu sebagai berikut:

1. menghitung *similaritas* menggunakan rumus *Jaccard coefficient* (Rahayu, 2009). Ukuran kemiripan antara objek ke$-i$dan objek ke $-j$ di hitung dengan rumusan

   $$si\ (X_i, X_j) = \frac{|X_i \cap X_j|}{|X_i \cup X_j|}, X_i \neq X_j$$

   Dimana :

   $si\ (X_i, X_j)$ = Ukuran kemiripan antara objek ke$-i$dan objek ke $-j$

   $i = 1,2,3, \dots, n\ d \qquad j = 1,2,3, \dots, n$

   X_i = himpunan pengamatan ke $-\ i$ dengan $X_i = \{ X_{1i}, X_{2i}, X_{3i,\dots,} X_{mk} \qquad i$

   X_j = himpunan pengamatan ke $-j$ dengan $X_j = \{ X_{1j}, X_{2j}, X_{3j,\dots,} X_{mk} \qquad j$

   |X| = bilangan kardinal atau jumlah anggota dari himpunan.

2. Langkah kedua adalah menentukan tetangga. Pengamatan dinyatakan sebagai tetangga jika nilai $si\ (X_i, X_j) \geq \Theta$.

3. Langkah terakhir adalah menghitung *link* antar objek pengamatan. Besarnya *link* dipengaruhi oleh nilai *threshold* (Θ) yang merupakan parameter yang ditentukan oleh pengguna yang dapat digunakan untuk mengontrol seberapa dekat hubungan antara objek . besarnya nilai Θ yang di inputkan adalah 0 < Θ < 1.

   Metode ROCK menggunakaan informasi tentang *link* sebagai ukuran kemiripan antar objek. Jika terdapat objek pengamatan X_i, X_j dan X_k dimana X_i tetangga dari X_j dan X_j tetangga dari X_k maka diikatakan X_i memiliki *link* dengan X_k walaupun X_i bukan tetangga dari X_k. Cara untuk menghitung *link* untuk semua kemungkinan pasangan dari *n* objek dapat menggunakan matriks A . matriks A merupakan matriks berukuran *nxn* yang bernilai 1 jika X_i dan X_j dinyatakan mirip (tetangga) dan bernilai 0 jika X_i dan X_j tidak mirip(bukan tetangga). Jumlah *link* antar pasangan X_i dan X_j di peroleh dari hasil kali antara baris ke $X_i i$ dan kolom ke X_j dari matriks A. Jika *link* antara X_i dan X_j semakin besar maka semakin besar pula kemungkinan X_i dan X_j berada dalam satu kelompok yang sama.

Algoritma *ROCK* yang didasarkan atas ukuran kebaikan (*goodness measure*) antar kelompok dengan rumusan pada persamaan *Goodness measure* adalah persamaan yang digunakan untuk menghitung jumlah *link* dibagi dengan kemungkinan *link* yang terbentuk berdasarkan ukuran kelompoknya (Tyagi &

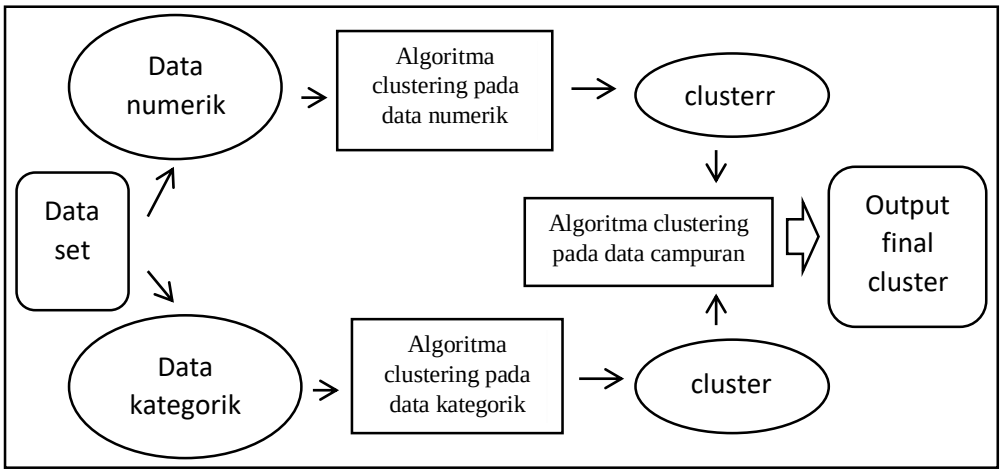
Sharma, 2012).

$$g\left(C_i, C_j\right) = \frac{li\ [C_i, C_j]}{(n_i, n_j)^{1+2f(\Theta)} - n_i{}^{1+2f(\Theta)} - n_j{}^{1+2f(\Theta)}}$$

Dengan $li\left[C_i, C_j\right] = \sum_{X_i \in C_i, X_j \in C_j} li\ \ (X_i\ , X_j)$ yang menyatakan jumlah *link* dari semua kemungkinan pasangan objek yang ada dalam C_i dan C_j serta n_i dengan n_j masing-masing menyatakan jumlah anggota dalam kelompok ke$-i$ dan j, sedangkan $f(\theta) = \frac{1-\theta}{1+\theta}$.

2.5 Pengelompokan Data Campuran

Salah satu medode yang bisa di gunakan untuk menyelesaikan masalah yang berkaitan dengan clustering data dengan tipe campuran (kategorik dan numerik) adalah metode ROCK . Pertama, data asli yang bertipe campuran dipisah menjadi dua yaitu data dengan tipe kategorik dan data dengan tipe numerik. Selanjutnya, kedua data tersebut diproses secara terpisah dengan menggunakan algoritma clustering yang sesuai dengan tipe masing-masing data. Terakhir, cluster-cluster yang dihasilkan oleh kedua algoritma digabungkan dan dipandang sebagai data baru dengan tipe kategorik, kemudian diproses dengan menggunakan algoritma clustering data kategorik untuk mendapatkan hasil akhir.Langkah dari pengelompokan data campuran ditunjukkan oleh Gambar 2.2 berikut:



Gambar 2.1 Langkah dari pengelompokan data campuran

Adapun langkah ensembel ROCK sebagai berikut :

1. Memisahkan data menjadi data kategorik dan data numerik
2. Melakukan clustering data numerik dengan menggunakan algoritma clustering data numerik dengan metode non hirarki.
3. Melakukan clustering data kategorik dengan menggunakan algoritma clustering data kategorik metode ROCK.
4. Menggabungkan output dari kedua algoritma tersebut menjadi data kategorik
5. Menggunakan kembali metode ROCK untuk melakukan clustering terhadap data hasil gabungan.

2.7 Kinerja Hasil *Clustering*

Kinerja Hasil *Clustering* Pengukuran kinerja hasil clustering merupakan langkah untuk mengetahui validitas suatu cluster. Cluster yang baik akan memiliki kehomogenan yang tinggi antar anggota dalam kelompok dan keheterogenan yang

tinggi antar kelompok (Hair, Black, Babin, & Anderson, 2010). Terdapat dua uji validitas yaitu validasi ukuran dan validasi metode.

- Validasi ukuran

Validasi ukuran yang digunakan dalam pemilihan jumlah cluster optimum pada variabel data berskala numerik adalah ukuran dunn index dan davie bouldin index. Dimana Dunn index adalah salah satu pengukuran validitas cluster yang diajukan oleh J.C.Dunn. Menurut Satato, Khotimah, & Muhammad (2015), validitas cluster berlandaskan pada fakta bahwa cluster yang terpisah pada umumnya memiliki jarak antar cluster yang besar dan jarak dalam cluster yang kecil. Dunn index tidak memiliki rentang nilai, nilai terbesar yang dihasilkan merupakan hasil ukuran terbaik terbaik (Dewanti, 2013).

$$D = min_{\,j=i+1..n_c} \left(min_{\,j=i+1..n_c} \frac{d\,[C_i, C_j]}{max_{\,k=1..n_c} (C_{\,k})} \right)$$

Sedangkan Davies bouldin index adalah salah satu metode evaluasi internal yang mengukur evaluasi cluster pada suatu metode pengelompokan yang didasarkan pada nilai kohesi dan separasi. Pengelompokan, kohesi dapat diartikan sebagai jumlah dari kedekatan data terhadap centroid dari cluster yang diikuti. Sedangkan separasi didasarkan pada jarak antar centroid dari clusternya. Semakin kecil nilai davies bouldin index maka semakin optimum jumlah cluster tersebut. Nilai Davies Bouldin Index (DBI) :

$$DBI = \frac{1}{K} \sum_{j=i}^{K} max \ \ (R_{i,j})$$

Dimana :

K = Jumlah cluster yang digunakan

$R_{i,j}$ = Jarak antara n cluster i dengan cluster j

- Validasi metode

Validasi metode yang di gunakan untuk menentukan metode pengelompokan data numerik terbaik dapat diketahui dari rasio nilai S_w dan S_B. Nilai perbandingan rasio S_w dan S_B ini juga di gunakan dalam memilih nilai *thresold* terbaik dari pengelompokan data kategorik dan data campuran.

Menurut Alvionita (2017), ukuran keragaman untuk data kategorik dikembangkan oleh Light dan Nargolin (1971), Okada (1999) serta Kader dan Perry (2007). Jika terdapat sebanyak *n* pengamatan dengan n_k merupakan jumlah pengamatan dengan kategori ke-*k* dimana *k* = 1,2,3,...K dan $\sum_{k=1}^{K} n_k = n$. Selanjutnya, n_k merupakan jumlah pengamatan dengan kategori ke-*k* dan kelompok ke-*c* , dimana c = 1,2,3,...C dengan C adalah jumlah yang terbentuk, sehingga $n_c = \sum_{k=1}^{K} n_k$ merupakan jumlah pengamatan pada kelompok ke-*c* $n_k = \sum_{c-1}^{K} n_k$ merupakan jumlah pengamatan pada kategori ke-*k*. Total jumlah pengamatan dapat di tuliskan menjadi

$$n = \sum_{c-1}^{K} n_c = \sum_{k=1}^{K} n_k = \sum_{c-1}^{K} \sum_{k=1}^{K} .$$

Jumlah kuadrat total atau SST pada sebuah peubah data kategorik dirumuskan sebagai berikut :

$$\text{SST} = \frac{n}{2} - \frac{1}{2n}\sum_{K=1}^{K} {n_k}^2$$

Rumus total jumlah kuadrat dalam kelompok atau SSW :

$$\text{SSW} = \sum_{c=1}^{C}\left(\frac{n_c}{2} - \frac{1}{2n_c}\sum_{k=1}^{K} {n_k}^2\right) = \frac{n}{2} - \frac{1}{2}\sum_{c=1}^{c}\frac{1}{n_c}\sum_{k}^{K} {n_k}^2$$

Rumus Jumlah kuadrat antar kelompok atau SSB :

$$\text{SSB} = \frac{1}{2}\sum_{c-1}^{c}\frac{1}{n_c}\sum_{k}^{K} {n_k}^2 \frac{1}{2n_c}\sum_{k}^{K} {n_k}^2$$

,

Rumus Mean of square (MST), mean of square (MSW), dan mean of square beetwen (MSB) sec ara bertutut-turut sebagai berikut:

$$\text{MST} = \frac{SST}{n-1}$$

$$\text{MSW} = \frac{SSW}{\text{n} - \text{C}}$$

$$\text{MSB} = \frac{SSB}{\text{C} - 1}$$

Rumus simpangan bakudalam kelompok(S_w) dan rumus simpangan baku antar kelompok (S_B) paada data kategorik adalah :

$$S_w = [MSW]^{1/2}$$

$$S_B = [MSB]^{1/2}$$

Sama halnya pada data numerik, untuk melihat kinerja suatu metode pengelompokan dapat di lihat dari nilai rasio antara (S_w) dan (S_B). Semakin kecil

nilai rasio maka semakin baik juga kinerja suatu metode. kinerja metode yang baik antinya terdapat homogenitas maksimum dalam cluster dan heterogenitas yang maksimum antar *cluster*.