BAB II
KAJIAN PUSTAKA DAN DASAR TEORI

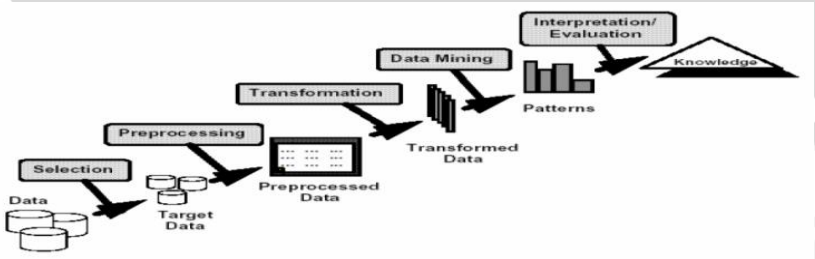
2.1 Kajian Pustaka

Penelitian sebelumnya yang pernah dilakukan oleh S. Aranganayagi dan K. Thangevel (2009) menggunakan *new dissimilarity measure* pada algoritma *K-Modes*. Pada penelitian itu ukuran jarak yang asli yang ada pada *K-Modes* diganti dengan ukuran jarak baru yaitu *new dissimilarity measure*. Kesimpulan yang didapat bahwa *new dissimilarity measure* menghasilkan kualitas *cluster* yang lebih baik dibanding *K-Modes* yang konvensional [ARA-09].

2.2 Data

Data sering disebut sebagai bahan mentah informasi. Data adalah kumpulan kejadian atau fakta-fakta yang dapat dijamin kebenarannya [Lee-10]. Tapi menurut Murdick, dkk (1984) merumuskan bahwa data adalah fakta yang tidak sedang digunakan pada proses keputusan, biasanya dicatat dan diarsipkan tanpa maksud untuk segera diambil kembali untuk pengambilan keputusan. Menurut Zulkifli Amsyah (1987) data adalah fakta yang sudah ditulis dalam bentuk catatan atau direkam ke dalam berbagai bentuk media.

Knowledge discovery in databases (KDD) adalah keseluruhan proses *non-trivial* untuk mencari dan mengidentifikasi pola (*pattern*) dalam data, dimana pola yang ditemukan bersifat sah, baru, dapat bermanfaat dan dapat dimengerti [USA-96]. Gambar 2.1 menunjukkan proses dari KDD.



Gambar 2.1 Proses *Knowledge discovery in databases* (KDD) [USA-96]

Berikut ini adalah proses dari KDD [KAN-03]:

1. *Data Selection*

   Menciptakan himpunan data target, pemilihan himpunan data, atau memfokuskan pada subset variabel atau sampel data, dimana penemuan dilakukan. 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*

   Pemrosesan pendahuluan dan pembersihan data merupakan operasi dasar seperti penghapusan *noise* dilakukan. 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, seperti kesalahan cetak (*tipografi*). Dilakukan proses *enrichment*, yaitu proses memperkaya data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan KDD, seperti data atau informasi eksternal.

3. *Transformation*

   Pencarian fitur-fitur yang berguna untuk mempresentasikan data bergantung kepada *goal* yang ingin dicapai. Merupakan proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*. Proses ini merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

4. *Data mining*

   Pemilihan tugas *data mining*, pemilihan *goal* dari proses KDD seperti klasifikasi, regresi, *clustering* dan lain-lain. Proses *data mining* yaitu proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, algoritma dalam *data mining* sangat bervariasi. Pemilihan algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

5. *Interpretation / Evaluation*

Penerjemahan pola-pola yang dihasilkan dari *data mining*. Pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya.

Data dapat dibagi menjadi 2 jenis yaitu data numerik dan data kategori.

2.2.1 Data Numerik

Data numerik adalah data yang berupa angka atau pecahan yang dapat diukur dan kontinyu. Data numerik dapat diukur karena nilai dari data numerik dapat dibandingkan mana yang lebih kecil dan yang lebih besar. Contoh dari data numerik adalah:

- Tinggi badan : 155, 160, 165, dan 170.
- Berat badan : 50, 58, 62, dan 70.
- Jumlah penduduk : 214, 125, 156, dan 179.

2.2.2 Data Kategori

Data kategori adalah data yang mengambil dari suatu himpunan nilai tertentu. Data kategori tidak harus berupa angka atau pecahan, misalkan warna mata, warna kulit, nama negara, dan merk produk. Data kategori berbeda dengan data numerik, data kategori tidak dapat diukur karena tidak dapat dibandingkan mana yang memiliki nilai lebih besar dan yang lebih kecil. Contoh dari data kategori adalah:

- Warna mata : hitam, coklat, dan biru.
- Warna kulit : hitam, sawo matang, kuning langsat, dan putih.
- Nama negara : Indonesia, Singapura, Malaysia, dan Jepang.
- Merk produk : aqua, swa, ades, dan vit.

2.3 Data Mining

2.3.1 Pengertian Data Mining

Data mining merupakan sebuah proses menganalisis sekumpulan data hasil penelitian, dengan tujuan untuk menemukan hubungan antar data, dan untuk meringkas data sehingga data menjadi mudah dimengerti dan berguna bagi pemiliki data [HAN-01].

Data mining didefinisikan sebagai satu set teknik yang digunakan secara otomatis untuk mengeksplorasi secara menyeluruh dan membawa ke permukaan relasi-relasi yang kompleks pada set data yang sangat besar [MOE-02].

Menurut Giri (2003), definisi sederhana dari *data mining* adalah ekstraksi informasi atau pola yang penting dan menarik dari data yang ada di *database* skala besar. Dalam jurnal ilmiah, *data mining* juga dikenal dengan nama *Knowledge Discovery in Databases* (*KDD*), dan menurut Pramudiono (2003), definisi umum dari *data mining* adalah serangkaian proses untuk menggali nilai tambah berupa pengetahuan yang selama ini tidak diketahui, secara manual dari suatu kumpulan data.

Sedangkan menurut Santosa (2007), *data mining* yang sering juga disebut *Knowledge Discovery in Databases* (*KDD*), adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar.

Ada beberapa definisi dari *data mining* menurut [KUS-07] antara lain yaitu :

1. *Data mining* adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual.
2. *Data mining* adalah analisa otomatis dari data yang berjumlah besar atau kompleks dengan tujuan untuk menemukan pola atau kecenderungan yang penting yang biasanya tidak disadari keberadaannya.
3. *Data mining* atau *Knowledge Discovery in Databases* (KDD) adalah pengambilan informasi yang tersembunyi, dimana informasi tersebut sebelumnya tidak dikenal dan berpotensi bermanfaat. Proses ini meliputi

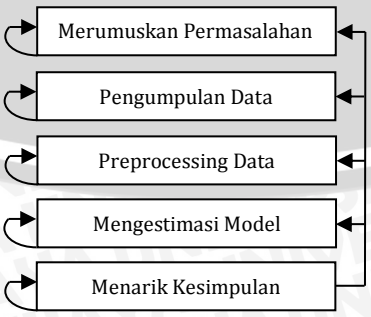
sejumlah pendekatan teknis yang berbeda, seperti *clustering*, data *summarization, learning classification rules*.

2.3.2 Proses Data Mining

Menurut Kantardzic (2003), dalam menemukan model dari data dalam data mining ada beberapa prosedur yang harus dilakukan, diantaranya:

1. Menentukan permasalahan dan mengformulasikan hipotesis.
   Pada praktiknya, proses ini digunakan untuk menyelaraskan hubungan antara ahli *data mining* dengan ahli aplikasi.
2. Pengumpulan Data.
   Proses ini berkonsentrasi pada bagaimana proses pembuatan data dan pengumpulan data.
3. Preprocessing Data.
   Terdapat dua sub fase dalam *preprocessing* data, yaitu:
   a. Data *Preparation* : Proses penyiapan data sebelum data diolah (dimodelkan).
   b. Data *Dimensionality reducation* : Proses peringkasan data.
4. Mengestimasi Model / Memperkirakan Model.
   Seleksi dan implementasi terhadap metode data mining yang tepat merupakan proses utama prosedur ini.
5. Menafsirkan model dan menarik kesimpulan.

Gambaran dari alur proses-proses yang terdapat pada *data mining*, diilustrasikan pada Gambar 2.2



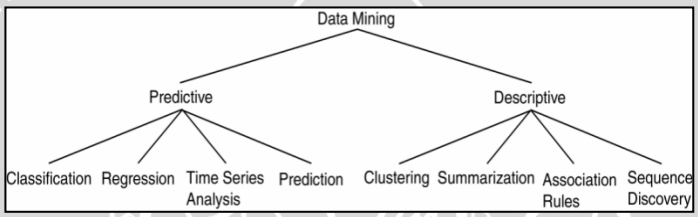
Gambar 2.2 Proses-proses dalam *data mining*

2.3.3 Metode Data Mining

Secara garis besar *data mining* dapat dikelompokkan menjadi 2 kategori utama, yaitu [TAN-05]:

1. *Descriptive mining*, yaitu proses untuk menemukan karakteristik penting dari data suatu basis data. Tehnik *data mining* yang termasuk dalam *descriptive mining* adalah *clustering*, *association*, dan *sequential mining*.
2. *Predictive*, yaitu proses untuk menemukan pola dari data dengan menggunakan beberapa variabel lain di masa depan. Salah satu tehnik yang terdapat dalam *predictive mining* adalah *klasifikasi*.

Pembagian tugas dan teknik *data mining* ditunjukkan oleh Gambar 2.3 [DUN-02].



Gambar 2.3 Model dan Teknik *Data mining*

2.3.4 Fungsi pada Data Mining

Menurut Larose (2005), pengelompokan *data mining* menurut fungsinya antara lain :

1. Fungsi deskripsi (*description*)

   Terkadang peneliti dan analis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecenderungan dalam data. Deskripsi dari pola dan kecenderungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecenderungan.
2. Fungsi estimasi (*estimation*)

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

3. Fungsi prediksi (*prediction*)

   Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil yang akan di masa mendatang.

4. Fungsi klasifikasi (*classification*)

   Merupakan tugas data mining yang mencari *mode*l yang mencari data baru yang belum pernah ada. Klasifikasi digunakan untuk prediksi data kategori, sedangkan untuk data numerik digunakan analisis regresi.

5. Fungsi peng*cluster*an (*clustering*)

   Peng*cluster*an merupakan pengelompokan *record*, pengamatan, atau memperhatikan dan membentuk *class* obyek-obyek yang memiliki kemiripan. *Cluster* adalah kumpulan *record* yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan dengan *record-record* dalam *Cluster* lain.

6. Fungsi asosiasi (*association*)

   Tugas asosiasi dalam *data mining* adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut *market basket analysis*.

2.3.4.1 Peng*cluster*an (*Clustering*)

Peng*cluster*an adalah suatu metode yang mengelompokkan sejumlah data besar menjadi suatu bagian-bagian kecil data yang mempunyai kemiripan letak, sifat, ciri, atau fitur lain yang telah ditentukan. Dalam mengelompokkan suatu data menjadi suatu bagian-bagian kecil diperlukan suatu titik yang akan dijadikan nilai parameter utama (*centroid of a group of dataset*) dari keseluruhan titik data yang terdapat pada kelompok.

Banyak manfaat dari *algoritma clustering* yang dapat kita terapkan dalam kehidupan sehari-hari seperti pengenalan bentuk (*pattern recognition*), kecerdasan buatan (*artificial intelligence*), *data mining*, dan lain-lain. Tipe-tipe *algoritma clustering* dapat dikelompokkan menjadi dua kategori, yaitu:

1. *Hierarchical methods*, misal : *Minimal Spanning Tree Method*.
2. *Nonhierarchical method*, misal : *K-Means Algorithm* dan *K-Modes Algorithm*.

Metode *clustering* yang baik akan menghasilkan *cluster* yang berkualitas tinggi yang bersifat :

1. Memiliki tingkat kemiripan yang tinggi dalam satu-*class similarity*.
2. Memiliki tingkat kemiripan yang tinggi dalam satu-*class dissimilarity*.

Kualitas *clustering* tergantung pada ukuran kemiripan yang digunakan oleh metode dan implementasinya. Kualitas *clustering* juga diukur kemampuannya untuk menemukan beberapa atau semua pola-pola yang tersembunyi.

2.4 Algoritma *K-Modes*

Dalam kehidupan nyata, data tidak hanya berupa numerik, banyak pula terdapat data kategori yang tidak dapat dijadikan numerik. Diperlukan suatu metode untuk mendapatkan informasi pada data kategori, salah satu metode tersebut adalah *K-Modes* yang diperkenalkan oleh Huang (1998).

Algoritma K-Modes melakukan dua modifikasi utama terhadap algoritma *K-means*, yaitu:

1. Cara mengukur jarak ketidaksamaan (*dissimilarity*).
2. Cara menentukan titik pusat.

Algoritma *K-Modes* adalah algoritma *clustering* yang diturunkan dari algoritma *K-Means*. Perbedaan *K-Modes* dengan *K-Means* adalah algoritma *K-Modes* untuk tipe data kategori sedangkan *K-Means* untuk tipe data numerik. Menurut Huang, algoritma *K-Means* memiliki kelemahan dimana penggunaan yang terbatas untuk data numerik. Huang melakukan modifikasi pada algoritma *K-Means*. Adapun modifikasi yang diaplikasikan adalah sebagai berikut :

1. Menggunakan *dissimilarity measure* untuk menghitung perbedaan atribut satu dengan lainnya.
2. Mengganti penggunaan rata-rata menjadi modus (nilai yang sering muncul).
3. Menggunakan metode berbasis frekuensi untuk melakukan pembaharuan terhadap modus.

Langkah-langkah dalam algoritma *K-Modes* adalah sebagai berikut.

1. Inisialisai pusat *cluster* sebanyak *k cluster*.
2. Bangkitkan *initial centroid* (titik pusat) secara acak dari data yang tersedia dimana banyaknya sesuai dengan *cluster* yang di bentuk (satu titik pusat yang di bentuk untuk tiap *cluster*)
3. Hitung jarak masing-masing data terhadap semua titik pusat *cluster*. Alokasikan setiap objek ke *cluster* terdekat menggunakan ukuran jarak *dissimilarity*.
4. Update pusat *cluster* menggunakan metode frekuensi.
5. Ulangi langkah 2 dan 3 sampai tidak ada objek yang berpindah *cluster* (*convergence*) untuk satu kali iterasi.

2.5 Similarity Measure, Dissimilarity Measure dan New Dissimilarity Measure

2.5.1 Similarity Measure

Menurut Han (2001), ukuran kesamaan (s*imilarity measure*) adalah kuantitas yang mencerminkan kekuatan hubungan antara dua obyek maupun dua fitur. Kuantitas memiliki kisaran 0 hingga 1. Jika ukuran kesamaan (s*imilarity measure*) antara fitur dua obyek telah dinotasikan, kuantitas dapat dihitung dengan beberapa cara tergantung skala pengukuran maupun tipe data yang dimiliki.

Dalam ukuran kesamaan (*similarity measure*), semakin besar nilainya berarti semakin mirip. Untuk ukuran kesamaan (*similarity measure*) dapat dipakai metode [SAN-07]:

1. *Cosinus*

   Cosinus antara dua titik x dan y didefinisikan sebagai

$$\cos\theta = \frac{x^T y}{\|x\|\|y\|} \quad (2.1)$$

   Dimana $\|x\|$ disefinisikan sebagai

$$\|x\| = \sqrt{\sum_{i=1}^{n} x_i^2} \quad (2.2)$$

2. Kovarian

   Kovarian antara dua data didefinisikan sebagai

$$cov\ (x,y) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) \quad 2.3$$

3. Korelasi

$$r\ (x,y) = \frac{cov\ (x,y)}{\sigma_x \sigma_y} \quad 2.4$$

2.5.2 Dissimilarity Measure

Ukuran ketidaksamaan (*dissimilarity measure*) adalah ukuran yang digunakan untuk menghitung jarak antar objek berkategori. Banyak metode dikembangkan yang memakai konsep jarak. Untuk mengukur jarak antara dua titik *x* dan *y*, (d(*x*,*y*)), kita bisa menggunakan beberapa konsep jarak. Ukuran jarak harus memenuhi syarat-syarat sebagai berikut [SAN-07]:

1. d(x,y) ≥ 0 (non-negatif)

   Tidak ada jarak yang mempunyai nilai negatif.

2. d(x,y) = 0 jika dan hanya jika x = y (*identity of indiscernibles*)

   Jarak antara suatu objek atau titik dengan objek atau titik itu sendiri adalah 0.

3. d(x,y) = d(x,y)

   Jarak dari x ke y adalah sama dengan jarak dari y ke x.

   d(x,y) ≤ d(x,y) + d(x,y) (ketidaksamaan (*dissimilarity*) segitiga)

Semakin kecil nilai ukuran jarak ketidaksamaan (*dissimilarity measure*) maka semakin dekat atau mirip antar objek tersebut.

Beberapa macam jarak yang sering digunakan dalam literatur data mining/*machine learning* adalah [SAN-07]:

1. Jarak *Euclidean* (L_2- *norm*)

   Jarak dua titik x dan y menurut *Euclidean* dirumuskan sebagai:

$$d(x,y) = \|x - y\|^2 = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad 2.5$$

2. Jarak *Manhattan* atau *Cityblock* (L_1- *norm*)

Menurut konsep ini, jarak dua titik x dan y dirumuskan:

$$d(x,y) = \sum_{i=1}^{n}(|x_i - y_i|) \quad 2.6$$

3. Jarak *Minkowski*

$$d(x,y) = \|x-y\|_q = \left(\sum_{i=1}^{n}(|x - y|^q)^{\frac{1}{q}}\right) \quad 2.7$$

Dimana q ≥ 1 adalah parameter yang bisa diseleksi. Dalam hal q=1, maka jarak tersebut menjadi jarak *Manhattan*. Sedangkan untuk q=2, jarak tersebut menjadi jarak Euclidean.

4. Jarak *Chebyshev* (L_∞- *norm*)

$$d(x,y) = \|x-y\|_\infty = \max_{1 \le i \le n} \{|x_i - y_i|\} \quad 2.8$$

5. Jarak *Mahalanobis* (*energy* - *norm*)

$$d(x,y) = \left[x - y^T \sum^{-1}(x - y)\right] \quad 2.9$$

Dimana $\sum$ adalah *covariance* matrix. Dalam hal $\sum$ = 1, maka jarak ini menjadi *Euclidean*.

Pada *K-Modes* yang asli ukuran jarak ketidaksamaan yang digunakan adalah *simple mismatching*. *Simple mismatching* dapat didefinisikan sebagai jumlah atribut yang tidak sama antara dua objek [HUA-98]. Misalkan *X* dan *Y* adalah dua objek berkategori dengan atribut sebanyak *m* maka persamaan *simple mismatching* dapat dilihat pada persamaan 2.10.

$$d(X,Y) = \sum_{j=1}^{m} \delta(x_j, y_j) \quad 2.10$$

dimana

$$\delta(x_j, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases} \quad 2.11$$

Pada persamaan 2.10 atau 2.11, x_j adalah atribut objek X yang ke j dan y_j adalah atribut objek Y yang ke j.

2.5.3 New Dissimilarity Measure

Sedangkan pada penelitian ini akan digunakan ukuran jarak ketidaksamaan baru (*new dissimilarity measure*) dengan merubah persamaan 2.11 dimana $x_j = y_j$ tidak lagi bernilai 0 tetapi 1-w_{ij}. Nilai w_{ij} adalah perkalian perbandingan nilai atribut di *cluster* dengan perbandingan nilai atribut di *dataset* seperti pada persamaan 2.12.

$$\delta(x_j, y_j) = \begin{cases} 1 - w_{ij} & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases} \quad 2.12$$

dimana

$$w_{ij} = \left(\frac{f\left(\frac{v_{ij}}{C_i}\right)}{|C_i|}\right) * \left(\frac{f\left(\frac{v_{ij}}{D}\right)}{|D|}\right) \quad 2.13$$

Keterangan:

$f(v_{ij}/C_i)$ adalah jumlah objek di *cluster* C_i dengan nilai atribut v_{ij}.

$f(v_{ij}/D)$ adalah jumlah objek di *dataset* dengan niali atribut v_{ij}.

C_i adalah jumlah objek di *cluster* C_i.

D adalah jumlah objek di *dataset*.

2.6 *Purity Measure*

Cluster dikatakan murni (*pure*) semua objek dengan *class* yang sama berada pada *cluster* yang sama. Untuk mengukur tingkat akurasi *clustering* atau '*r*', Huang (1998) mengusulkan cara mengukurnya menggunakan persamaan 2.14.

$$r = \frac{1}{n}\sum_{i=1}^{k} a_i \qquad 2.14$$

Dimana :

r adalah tingkat akurasi *klustering*

k adalah jumlah c*luster*

a_i adalah objek yang muncul di dalam *cluster* Ci dan pada label *class* yang sesuai.

Sedangkan untuk menghitung *error cluster* atau '*e*' seperti pada persamaan 2.15.

$$e = 1 - r \qquad 2.15$$

Dimana *r* adalah nilai tingkat kemurnian *cluster*. Misalkan ada *dataset* dengan label *class* D1 ada 10 objek, D2 ada 10 objek, D3 ada 10 objek, dan D4 ada 17 objek. Setelah *dataset* di*cluster*kan menjadi 4 *cluster* hasilnya seperti pada tabel 2.1.

Tabel 2.1 Tabel Contoh Hasil *Clustering*

Class	***Cluster* 1**	***Cluster* 2**	***Cluster* 3**	***Cluster* 4**
D1	-	-	-	10
D2	-	-	10	-
D3	10	-	-	-
D4	2	15	-	-

Dilihat dari tabel 2.1 pada kluster 1 terdapat 10 objek D3 dan 2 objek D4 selain itu *class* D4 juga ada pada *cluster* 2 sebanyak 15 objek. Jadi nilai tingkat kemurnian (*Purity Measure*) dan

- Tingkat Kemurnian (*Purity Measure*)

  $r = \frac{10+15+10+10}{47} = 0.9574$

- *Error Clustering*

  $e = 1 - 0.9574 = 0.0426$

2.7 Pengujian *Accuracy*

Metode evaluasi *Performance Metric Error rate* digunakan untuk mengevaluasi nilai atribut yang bertipe kualitatif. Menurut Tan.,dkk. (2004), *Error rate* dapat dihitung dengan persamaan 2.16.

$$Error\ rate = \frac{Jumlah\ Prediksi\ Salah}{Jumlah\ Total\ Prediksi} \quad 2.16$$

Sedangkan *accuracy* adalah kebalikan dari *error rate*. *Accuracy* dapat dihitung dengan persamaan 2.17.

$$Accuracy = 1 - Error\ Rate \quad 2.17$$

2.8 Pengujian F-Measure

Pada proses peng*cluster*an, prosedur untuk megevaluasi hasilnya dikenal sebagai validitas *cluster* dan dapat dilakukan dengan berbagai macam pengukuran yang disebut dengan pengukuran validitas *cluster* [PET-00].

Pengukuran validitas pada penelitian ini menggunakan *F-Measure* karena sudah ada data pembanding eksternal yang dapat digunakan dalam evaluasi. Untuk menguji *accuracy* dari hasil *F-Measure*, digunakan suatu standar yang disebut matriks *confusion*. Matriks *confusion* berisi informasi mengenai peng*cluster*an yang sebenarnya dan prediksi peng*cluster*an yang dilakukan oleh sistem [HAM-03]. Tabel 2.2 menunjukkan matriks *confusion* [LEW-95].

Tabel 2.2 Matriks *Confusion*

		Peng*cluster*an Sebenarnya	
		***Cluster* Sama**	***Cluster* Beda**
Peng*cluster*an Sistem	***Class* Sama**	*True Positive*	*False Negative*
	***Class* Beda**	*False Positive*	*True Negative*

1. *True Positive (TP)* menunjukkan bahwa *record* yang termasuk dalam hasil peng*cluster*an oleh sistem memang merupakan anggota *cluster*. Peng*cluster*an sebenarnya ya dan peng*cluster*an oleh sistem ya.
2. *False Negative (FN)* menunjukkan bahwa *record* yang termasuk dalam hasil peng*cluster*an oleh sistem ternyata seharusnya bukan merupakan anggota *cluster*. Peng*cluster*an sebenarnya bukan dan peng*cluster*an oleh sistem ya.
3. *False Positive (FP)* menunjukkan bahwa *record* yang tidak termasuk dalam hasil peng*cluster*an oleh sistem ternyata seharusnya merupakan anggota *cluster*. Peng*cluster*an sebenarnya ya dan peng*cluster*an oleh sistem bukan.
4. *True Negative (TN)* menunjukkan bahwa *record* yang tidak termasuk dalam hasil peng*cluster*an oleh sistem ternyata seharusnya bukan merupakan anggota *cluster*.

Ada beberapa standar pengukuran yang digunakan dalam peng*cluster*an, diantaranya adalah *recall*, *precision* dan *F-measure*. *Recall* adalah tingkat keberhasilan mengenali suatu *event* dari seluruh *event* yang seharusnya dikenali. Dari pengertian tersebut, *recall* berarti ukuran dari jumlah *record* benar yang berhasil dikelompokan oleh sistem dari keseluruhan *record* yang seharusnya benar [TAL-03]. Cara mencari nilai *recall* ini adalah dengan membagi jumlah *record* hasil peng*cluster*an sistem yang benar dengan jumlah *record* yang sebenarnya dalam *cluster* tersebut.

$$recall = \frac{TP}{TP + FP} \qquad 2.18$$

Sedangkan *precision* adalah tingkat ketepatan hasil pengelompokan terhadap suatu *event*. Artinya, dari seluruh *record* hasil peng*cluster*an, berapa persenkah yang dinyatakan benar [TAL-03].

$$precision = \frac{TP}{TP} \qquad 2.19$$

F-measure merupakan gabungan antara *recall* dan *precision y*ang didefinisikan dengan persamaan 2.20 [YAN-99]:

$$F = \frac{2 \text{ x } recall \text{ x } precision}{recall + precision} \qquad 2.20$$