Materi 4

DATA MINING

3 SKS | Semester 6 | S1 Sistem Informasi | UNIKOM | 2016

Nizar Rabbi Radliya | nizar.radliya@yahoo.com

Nama Mahasiswa	
NIM	
Kelas	

Kompetensi Dasar

Mampu melakukan perhitungan similaritas dan dissimilaritas objek satu atribut dan multiatribut.

Pokok Bahasan

Similaritas dan Dissimilaritas:
1. Ketidakmiripan dan kemiripan data satu atribut
2. Ketidakmiripan data multiatribut
3. Kemiripan data multiatribut

Kemiripan (*similarity*) adalah ukuran numerik dimana dua objeknya mirip, nilai 0 jika tidak mirip dan nilai 1 jika mirip penuh. Sementara ketidakmiripan (*dissimilarity*) adalah derajat numerik dimana dua objek yang berbeda, jangkauan nilai 0 sampai 1 atau bahkan sampai ∞.

I. Ketidakmiripan dan Kemiripan Data Satu Atribut

Istilah ketidakmiripan juga dapat disebut sebagai ukuran jarak (*distance*) antara dua data. Jika s adalah ukuran kemiripan dan d adalah ukuran ketidakmiripan, serta jika interval/range nilainya adalah [0,1], maka dapat dirumuskan bahwa s+d=1. Sebenarnya ukuran kemiripan dan ketidakmiripan tidak harus selalu dalam interval [0,1], tetapi boleh juga menggunakan interval seperti [0,10], [0,100], bahkan menggunakan nilai negative seperti [-1,1], [-10,10] dan sebagainya. Transformasi nilai s dan d tidak hanya terbatas pada formula s+d=1, karena ada juga yang menggunakan s = $\frac{1}{1+d}$ atau s = e^{-d}.

Pada metode tertentu dalam klasifikasi, ada juga yang mengharuskan agar nilai interval ketidakmiripan data harus ditransformasi dalam interval yang ternormalisasi [0,1]. Sebagai contoh, ada data dengan nilai ketidakmiripan {10, 12, 25, 30, 40} dengan intervalnya [10,40]. Jika akan ditransformasi ke dalam interval [0,1], kita bisa menggunakan formula x = $\frac{x-\min(x)}{\max(x)-min(x)}$ sehingga nilai-nilai ketidakmiripan tersebut ditransformasi menjadi {0, 0.667, 0.5, 0.6667, 1}.

Untuk fitur yang menggunakan tipe ordinal, misalnya sebuah atribut yang mengukur kualitas produk dengan skala {rusak, jelek, sedang, bagus, sempurna}, skala tersebut harus ditransformasikan ke dalam nilai numerik, misalnya {rusak=0, jelek=1, sedang=2, bagus=3, sempurna=4}. Kemudian, ada dua produk P1 dengan kualitas bagus dan P2 dengan kualitas jelek. Jarak (ketidakmiripan) antara P1 dan P2 dapat dihitung dengan cara D(P1,P2) = 3-1 = 2, atau jika dalam interval [0,1] menjadi $\frac{3-1}{4}$ = 0.5, sedangkan nilai kemiripannya adalah 1-0.5 = 0.5.

Untuk atribut bertipe numerik (interval dan rasio), nilai ketidakmiripan didapat dari selisih absolut di antara dua data. Misalnya atribut usia, jika P1 adalah usia 45 dan P2 usia 25, sedangkan jangkauan nilai usia dalam data adalah [5,75], nilai ketidakmiripan P1 dan P2 adalah D(P1,P2) = 45-25 = 20, atau jika dalam interval [0,1] menjadi $\frac{20-5}{75-5}$ = 0.21, sedangkan nilai kemiripannya adalah 1-0.21 = 0.79.

Tabel 1. Formula Kemiripan dan Ketidakmiripan dengan Satu Atribut

Tipe Atribut	Kemiripan	Ketidakmiripan
Nominal	s = $\begin{cases} 1 \; jika \; x = y \\ 0 \; jika \; x \neq y \end{cases}$	d = $\begin{cases} 0 \; jika \; x = y \\ 1 \; jika \; x \neq y \end{cases}$
Ordinal	s = 1 – d	d = |x – y|/(n – 1)
Interval dan Rasio	s = -d, s = $\frac{1}{1+d}$, s = e^{-d}, s = $\frac{d-\min(d)}{\max(d)-min(d)}$	D = |x – y|

II. Ketidakmiripan Data Multiatribut

Terdapat banyak cara yang dapat digunakan untuk menghitung jarak (ketidakmiripan) dua objek dengan beberapa atribut, diantaranya:

1. Jarak Euclidian

$$D(x,y) = \sqrt{\sum_{j=1}^{n} |x - y|^2}$$

2. Jarak Manhattan/City Block

$$D(x,y) = \sum_{j=1}^{n} |x - y|$$

3. Jarak Chebyshev

$$D(x,y) = \max_{j=1} N(|x_j - y_j|)$$

Sebagai contoh kita akan melakukan pengukuran jarak antar objek dengan persamaan euclidean pada set data tabel 2 di bawah ini.

Tabel 2. Contoh Data Dua Dimensi

Point	x	y
P1	1	1
P2	4	1
P3	1	2

Tabel 3. Hasil Pengukuran Jarak Euclidean

Euclidean	P1	P2	P3
P1	0	3	1
P2	3	0	3.16
P3	1	3.16	0

III. Kemiripan Data Multiatribut

Terdapat beberapa cara yang dapat digunakan untuk menghitung kemiripan (similaritas) dua objek dengan beberapa atribut, diantaranya:

1. Simple Matching (SMC) dan Jaccard Coefficients (J)

   Metode yang dapat digunakan untuk menghitung kemiripan dua objek (vektor) biner. Misalkan kita akan menghitung kemiripan objek p dan q, dimana kedua objek tersebut terdiri dari beberapa atribut yang bernilai biner. Kemiripan antara dua objek tersebut dapat dihitung dengan kuantitas berikut:

   M01 = Jumlah atribut dimana p adalah 0 dan q adalah 1

   M10 = Jumlah atribut dimana p adalah 1 dan q adalah 0

   M00 = Jumlah atribut dimana p adalah 0 dan q adalah 0

   M11 = Jumlah atribut dimana p adalah 1 dan q adalah 1

   Nilai dari M01, M10, M00 dan M11 dimasukan ke dalam persamaan Simple Matching (SMC) atau Jaccard Coefficients (J) di bawah ini:

   SMC = number of matches / number of attributes

   SMC = (M11 + M00) / (M01 + M10 + M11 + M00)

   J = number of 11 matches / number of not-both-zero attributes values

   J = (M11) / (M01 + M10 + M11)

2. Cosine Similarity

   Metode ini digunakan untuk menghitung dua obje (vektor) dokumen. Jika d1 dan d2 adalah dua vektor dokumen maka kemiripan antara dua vektor dokumen yang dihitung dengan cosine similarity adalah sebagai berikut:

   Cos(d2,d2) = (d1.d2) / | |d1| | | |d2| |

3. Extended Jaccard Coefficient (Tanimoto)

   Merupakan pengembangan dari Jaccard Coefficient untuk menghitung kemiripan dua objek (vektor) dengan atribut kontinyu, dengan persamaan sebagai berikut:

   $T(p, q) = \frac{p \cdot q}{||p||^2 + ||q||^2 - p.q}$

IV. Kemiripan Objek dengan Atribut Campuran

Terkadang ditemukan bahwa atribut (fitur) pada objek (vektor) terdiri dari tipe data campuran (tidak seragam). Tidak selalu semua atribut bertipe numerik (interval atau rasio) atau bertipe kategoris (nominal dan ordinal). Untuk menghitung nilai kemiripan objek dengan atribut campuran kita bisa gunakan persamaan berikut:

$s(x, y) = \frac{\sum_{i=1}^{r} S_i(x,y)}{\sum_{i=1}^{r} w_i}$

s_i merupakan ukuran kemiripan diantara fitur ke-I dari x dan y, sedangkan w_i adalah faktor bobot yang berkorelasi dengan fitur ke-1. Cara menentukan faktor bobot adalah dengan ketentukan sebagai berikut:

$w_i = 0$, apabila nilai fitur-I dari objek x dan y tidak teridentifikasi.

$w_i = 0$, apabila nilai $s_i = 0$.

$w_i = 1$, apabila selain dua syarat di atas.

V. Daftar Pustaka

[1] Astuti, F.A. 2013. Data Mining. Yogyakarta: Andi.
[2] Kusrini & Taufiz, E.L. 2009. Algoritma Data Mining. Yogyakarta: Andi.
[3] Prasetyo, E. 2012. Data Mining: Konsep dan Aplikasi Menggunakan MATLAB. Yogyakarta: Andi.
[4] Prasetyo, E. 2014. Data Mining: Mengolah Data Menjadi Informasi Menggunakan MATLAB. Yogyakarta: Andi.

VI. Materi Berikutnya

Pokok Bahasan	Klasifikasi
Sub Pokok Bahasan	1. Konsep klasifikasi 2. Klasifikasi berbasis *decision tree* 3. Pembahasan algoritma