Kesamaan Data Biner Berdasarkan Kategori Nilai Entropy dan Pola Struktur

Similarity for Binary Data Based on the Value of Entropy and Structure Patterns Categories

Kariyam
Departments of Statistics, Faculty of Mathematics and Natural Sciences
Islamic University of Indonesia

ABSTRACT

Similarity of two objects that have a form of binary data, usually calculated based on the frequencies in the contingency table that includes all discrete random variables. In this article we will discuss the similarity measures for binary data based on entropy values and structural patterns of the two object categories. Measuring similarity based on the value of entropy and structural pattern of categories can be used as a validation measure of similarity for binary data.

Keywords: Similarity, binary data, entropy, structure of patterns categories

PENDAHULUAN

Teknik pengelompokan untuk data biner sangat berbeda dengan data numerik dalam hal ukuran similaritas atau kemiripan. Data biner nominal merupakan tipe data kategorik yang hanya menggunakan dua kategori, yang biasanya sering diinisialkan sebagai 0 dan 1. Misalnya variabel jenis kelamin, dimana 0 untuk laki-laki, dan 1 untuk perempuan, atau variabel hasil kelulusan, dimana 0 untuk tidak lulus, dan 1 untuk lulus.

Apabila seluruh variabel yang dimiliki oleh dua buah obyek mempunyai data bertipe biner, maka umumnya ukuran similaritas antara dua obyek tersebut didefinisikan berdasarkan frekuensi data dalam tabel kontingensi pada nilai yang sama (*matches*) dan nilai yang tidak sama (*mismatches*) untuk semua p variabel. Kajian tentang ukuran similaritas untuk data biner, seperti dikutip oleh Everitt *et al*. (2001) di halaman 38, telah banyak dilakukan, diantaranya yaitu ukuran similaritas koefisien Jaccard (1908), Rogers & Tanimoto (1960), Sokal & Sneath (1963), serta Gower & Legendre (1986).

Misalkan dua obyek i dan j masing-masing diamati pada p variabel random diskret bertipe biner, maka tabel kontingensi dapat disajikan sebagaimana tabel 1. Pada tabel 1., nilai a dan nilai d, menunjukkan frekuensi data yang sama (*matches*), yaitu baik obyek i maupun obyek j, mempunyai kategori 0 (nol) sebanyak a, dan mempunyai kategori 1 (satu) sebanyak d. Sebaliknya, nilai b dan nilai c, menunjukkan frekuensi data yang tidak sama (*mismatches*). Secara sederhana, jika frekuensi a dan frekuensi d dijumlahkan hasilnya mendekati jumlah seluruh variabel (p), maka obyek i dan obyek j, dikatakan semakin mirip. Apabila $a + d = p$, maka obyek i dan obyek j, dikatakan identik.

Tabel 1. Tabel kontingensi data biner pada dua obyek

Hasil		Obyek *i*		Jumlah
		1	0	
Obyek *j*	1	a	b	$a + b$
	0	c	d	$c + d$
Jumlah		$a + c$	$b + d$	$p = a + b + c + d$

Beberapa ukuran similaritas antara obyek i dengan obyek j (S_{ij}), yang telah diusulkan diantaranya :

Ukuran similaritas Jaccard :

$$S_{ij} = \frac{a}{a+b+c} \quad (1)$$

Ukuran similaritas Rogers – Tanimoto :

$$S_{ij} = \frac{a+d}{[a+2(b+c)+d]} \quad (2)$$

Ukuran similaritas Sokal – Sneath :

$$S_{ij} = \frac{a}{[a+2(b+c)]} \quad (3)$$

Ukuran similaritas Gower – Legendre :

$$S_{ij} = \frac{a+d}{\left[a+\frac{1}{2}(b+c)+d\right]} \quad (4)$$

Ukuran similaritas pada persamaan (1), (2), (3), dan (4), dihitung berdasarkan frekuensi sel di tabel kontingensi (tabel 1.) dengan bobot yang bervariasi. Kelebihan dari ukuran-ukuran similaritas di atas adalah cara perhitungannya yang sederhana dan mudah. Ukuran similaritas tersebut juga hanya sesuai untuk data bertipe biner, dan kurang sesuai jika diterapkan untuk data non metrik dengan *level* lebih dari dua kategori. Secara prinsip perhitungan ukuran similaritas antar obyek ini merupakan langkah awal dalam proses pengelompokan obyek.

Perluasan ukuran similaritas data biner untuk data non metrik lebih dari dua *level* juga telah dibahas oleh Everitt, *et al*. (2001) di halaman 38 – 39. Misalkan dua obyek i dan j masing-masing mempunyai p variabel dengan *level* lebih dari dua, katakan l *level,* maka nilai similaritas kedua obyek merupakan rata-rata dari koefisien similaritas p variabel, sebagai berikut:

$$S_{ij} = \frac{1}{p}\sum_{k=1}^{p} S_{ijk} \quad (5)$$

Nilai similaritas S_{ijk} yang dihitung dari persamaan (2) dan (4), akan bernilai 1 (satu) ketika dua obyek memiliki kategori sama (misalkan 4 dengan 4, 1 dengan 1, dst), dan akan bernilai nilai 0 (nol) yaitu ketika kedua obyek memiliki kategori tidak sama. Misalkan u ($u \le p$) adalah jumlah kategori dari variabel random diskret yang bernilai sama pada dua obyek, maka persamaan (5) yang dihitung dengan menggunakan persamaan (2), dan persamaan (4), dapat dituliskan sebagai:

$$S_{ij} = \frac{u}{p} \quad (6)$$

Menurut penulis, persamaan (6) mempunyai kelemahan, yaitu bahwa ukuran similaritas tersebut tidak mempertimbangkan **struktur perbedaan** kategori dari setiap variabel. Artinya ketika pada variabel ke – k , obyek i mempunyai kategori 1, dan obyek j mempunyai kategori 4, maka dipandang **sama** dan bernilai nol, dengan ketika variabel ke – k , obyek i mempunyai kategori 3, dan obyek j mempunyai kategori 4. Dengan demikian kombinasi kategori dua obyek berapa saja, selalu dipandang sama sebagai data biner.

Kajian tentang penggunaan nilai *entropy* klasikal dalam proses pengelompokan, telah dilakukan oleh Chen K & Liu L, (2005). Misalkan dipunyai himpunan data X, yang terdiri dari N pengamatan pada p variabel random diskret $X = (x_1, x_2, ..., x_p)$. Kategori setiap komponen x_k, $1 \le k \le p$ diambil dari domain A_k, yang berhingga. Misalkan juga $p(x_k = v);\ v \in A_k$ menunjukkan probabilitas dari $x_k = v$ dalam himpunan data X yang memuat N pengamatan. Selanjutnya Chen dan Liu, dalam makalahnya di halaman 2, mendefinisikan nilai *entropy* dalam suatu himpunan data X ($H(X)$) yang terdiri dari N pengamatan dan p variabel tersebut, sebagai berikut:

$$H(X) = -\sum_{k=1}^{p}\sum_{v\in A_k} p(x_k = v)\log_2 p(x_k = v) \quad (7)$$

Menurut penulis, persamaan (7) ini mempunyai kekurangan, yaitu **belum** secara langsung menggunakan struktur pola kategori penyusun kelompok. Demikian juga, ketika himpunan data hanya berisi dua obyek, rumusan *entrophy* ini kurang lazim digunakan sebagai ukuran similaritas antara kedua obyek tersebut.

Improvisasi algoritma pengelompokan data kategori dengan berdasarkan pada total kemungkinan kombinasi yang terjadi dari p variabel secara serentak, telah diusulkan oleh Deng S, Xu X & He Z, (2006). Misalkan $A_1, A_2, \ldots, A_p$ adalah himpunan atribut kategori dengan domain $D_1, D_2, \ldots, D_p$. Misalkan pula himpunan data D menunjukkan himpunan obyek dari setiap obyek $t:\ t \in D_1\ x\ D_2\ x\ ...x\ D_p$. Selanjutnya $VAL_1, \ldots, VAL_p$ menotasikan himpunan atribut yang berbeda untuk nilai-nilai $A_1, A_2, \ldots, A_p$. Untuk setiap $v_{ij} \in VAL_i$, $f(v_{ij})$ menotasikan frekuensi dari kejadian yang mungkin dalam D. Selanjutnya Deng, *et al*. (2006), di halaman 24, mendefinisikan *More Similar Attribute Value Set (MSFVS)* sebagai berikut :

$$MSFVS(v_{ij}) = \left\{ v_{ik} \,\middle|\, f(v_{ik}) \le f(v_{ij}) \text{ dan } v_{ik} \in VAL_i \right\} \quad (8)$$

Dengan mengambil n sebagai total banyaknya obyek dalam himpunan data, maka bobot dari nilai atribut v_{ij} didefinisikan sebagai berikut:

$$W(v_{ij}) = 1 - \sum_{v_{ik} \in MSFVS} \frac{f(v_{ik})(f(v_{ik})-1)}{n(n-1)} \quad (9)$$

Selanjutnya persamaan (9) ini digunakan sebagai dasar untuk menghitung similaritas antar kelompok. Ide dasar dari kajian ini hampir sama dengan nilai *entrophy,* yaitu menggunakan frekuensi terjadinya nilai atribut dalam suatu kelompok, untuk bahan perhitungan pembentukan jumlah dan anggota kelompok. Namun sebagaimana pada *entropy,* persamaan (9) ini kurang lazim digunakan untuk $n = 2$ buah (data biner).

Widodo E, Guritno S, Haryatmi S & Kariyam (2009), telah melakukan kajian tentang ukuran similaritas data ordinal dengan mempertimbangkan secara langsung struktur pola kategori obyek penyusun kelompok. Ukuran similaritas data ordinal yang diusulkan didasarkan pada konteks permasalahan penelitian. Artinya ketika dihadapkan pada himpunan data non metrik, khususnya tipe ordinal, misalkan pada konteks permasalahan, angka 4 untuk kategori jawaban sangat baik, angka 3 untuk kategori jawaban baik, angka 2 untuk kategori jawaban kurang baik, dan angka 1 untuk kategori jawaban sangat tidak baik, maka struktur pola kategori 3 dan 4 akan **dianggap berbeda** dengan struktur pola kategori 1 dan 4. Dalam kajiannya telah dikembangkan dan diusulkan ukuran similaritas untuk data ordinal dengan mempertimbangkan struktur pola kategori sesuai konteks permasalahan semula. Misalkan dipunyai himpunan data *S* dengan N pengamatan dan p variabel random diskret $\mathbf{X} = (x_1, x_2, ..., x_p)$. Untuk setiap variabel random diskret x_k, $1 \le k \le p$, mempunyai nilai yang diambil dari domain A_k berhingga dimana banyaknya kategori dalam domain A_k adalah v_k, dan kategori A_k berbeda dari A_l untuk suatu $(k \ne l)$. Widodo *et al*. (2009) mengusulkan ukuran similaritas dengan terlebih dahulu menentukan bobot struktur pola kategori dalam variabel antara dua obyek. Langkah pertama penentuan bobot struktur pola angka, dimulai dengan memberikan rangking (r_k), pada kategori setiap variabel.

Jika variabel ke – k mempunyai v_k kategori, maka $r_k \in \{1, 2, ..., v_k\}$. Langkah kedua adalah menghitung selisih positif rangking dari struktur pola kategori yang terbentuk pada dua obyek (w). Misalkan nilai obyek i pada variabel ke – k setelah dirangking adalah r_k^i dan nilai obyek j pada variabel ke – k setelah dirangking adalah r_k^j, maka nilai w dihitung sebagai berikut:

$$w = \left| r_k^i - r_k^j \right| \quad (10)$$

Langkah terakhir, bobot struktur pola kategori, dinotasikan dengan b_k, yang terbentuk dari dua obyek i dan obyek j pada variabel ke – k, dihitung sebagai berikut:

$$b_k = \frac{w}{v_k - 1} \quad (11)$$

Secara sederhana, Widodo *et al*. (2009), mendefinisikan ukuran similaritas antara dua obyek i dan obyek j, (S_{ij}), sebagai berikut:

$$S_{ij} = \sum_{k=1}^{p} b_k \quad (12)$$

Berdasarkan hasil-hasil kajian di atas, maka pada tulisan ini akan dibahas perbandingan **aplikasi** beberapa ukuran similaritas data biner. Ukuran similaritas yang dibandingkan, diambilkan secara langsung dari sejumlah ukuran similaritas data biner yang sudah ada, dan ukuran similaritas yang diturunkan dari data ordinal. Tulisan ini akan dibatasi pada perbandingan aplikasi ukuran similaritas data biner yang diusulkan oleh Rogers – Tanimoto, Gower – Legendre, Widodo, dkk, dan nilai *entropy.* Hasil perbandingan yang diperoleh, diharapkan dapat memberikan alternatif validasi pemecahan masalah ukuran similaritas data biner.

HASIL DAN PEMBAHASAN

Ukuran similaritas data biner berbasiskan pada nilai *entropy*

Misalkan dipunyai himpunan data **sampel *S*** dengan n pengamatan dan p variabel random diskret, yaitu himpunan sampel dari vektor random diskret $\mathbf{X} = (x_1, x_2, ..., x_p)$. Untuk setiap komponen x_k, $1 \le k \le p$ nilai-nilainya diambil dari domain A_k berhingga yang berbeda dengan A_l $(k \ne l)$. Misalkan

$p\left(x_k = v\right);\ v \in A_k$ menunjukkan probabilitas dari $x_k = v$ dalam himpunan data sampel S. Estimasi nilai *entropy* $H(X)$ pada persamaan (7) dengan menggunakan himpunan data **sampel** ***S***, dituliskan sebagai $\hat{H}(X) = H\left(X \mid S\right)$, yaitu :

$$H(X|S) = -\sum_{k=1}^{p}\sum_{v\in A_k} p(x_k = v \mid S)\log_2 p(x_k = v \mid S) \quad (13)$$

Misalkan himpunan data sampel S dipartisi dalam ***G*** kelompok, yaitu $C^G = \{C_1, C_2, ..., C_g\}$ dengan n_g menunjukkan jumlah anggota obyek dalam kelompok C_g. Misalkan pula bahwa $p\left(x_k = v\right);\ v \in A_k$ sebagai probabilitas dari $x_k = v$ dalam kelompok C_g yang berisi n_g anggota. Nilai *entropy* untuk suatu kelompok C_g yang mempunyai anggota kelompok sejumlah n_g adalah:

$$H(X|C_g) = -\sum_{k=1}^{p}\sum_{v\in A_k} p(x_k = v \mid C_g)\log_2 p(x_k = v \mid C_g) \quad (14)$$

Pada kasus suatu kelompok mempunyai anggota (obyek) n_g = 2, dan p variabel random diskret yang diamati pada kedua obyek tersebut bertipe biner (dinotasikan 0 dan 1), dan dengan mengambil $p\left(x_k = 0 \mid C_g\right) = p_k$, serta $p\left(x_k = 1 \mid C_2\right) = q_k$, maka nilai *entropy* untuk persamaan (14), dapat dituliskan sebagai berikut:

$$H(X|C_g) = -\sum_{k=1}^{p}\sum_{v=0}^{1} p\left(x_k = v \mid C_g\right)\log_2 p\left(x_k = v \mid C_g\right)$$
$$= -\sum_{k=1}^{p}\left(p_k \log_2 p_k + q_k \log_2 q_k\right) \quad (15)$$

Apabila kedua obyek mempunyai kategori identik pada semua, *p,* variabel, maka nilai entropinya akan sama dengan nol. Sedangkan apabila terdapat *u* variabel $(u < p)$ yang kategorinya identik di kedua obyek yaitu keduanya 0 (nol) atau keduanya 1 (satu); dan *m* variabel $(m = p - u)$ dengan kategori berbeda di kedua obyek yaitu salah satu obyek mempunyai kategori nol atau satu, maka nilai *entropy* kedua obyek dapat dituliskan dengan :

$$H\left(X \mid C_g\right) = -m\left(\frac{1}{2}.\log_2\left(\frac{1}{2}\right)\right) = m \quad (16)$$

Ukuran similaritas data biner berbasiskan pada struktur pola kategori

Misalkan dipunyai himpunan data dengan N pengamatan dan p variabel random diskret $X = \left(x_1, x_2, ..., x_p\right)$ bertipe biner. Menurut Widodo, dkk, (2009), maka struktur pola kategori antara dua buah obyek i dan obyek j, yang mungkin yaitu $\{\ (0,0);\ (0,1);\ (1,1);\ (1,0)\ \}$. Sedangkan nilai w yang mungkin pada persamaan (10) yaitu 0 dan 1, dan bobot dari setiap struktur pola kategori b_k untuk $\{\ (0,0);\ (0,1);\ (1,1);\ (1,0)\ \}$ menurut persamaan (11) masing-masing adalah $\{\ 0;\ 1;\ 0;\ 1\ \}$. Apabila dua buah obyek i dan obyek j, diamati pada p variabel bertipe biner, dan ditemukan u $(u \le p)$ buah variabel yang mempunyai kategori sama di kedua obyek, dan m variabel $(m = p - u)$ dengan kategori berbeda di kedua obyek yaitu salah satu obyek mempunyai kategori nol atau satu, maka ukuran similaritas pada persamaan (12) menurut usulan Widodo *et al*. (2009), menjadi sangat sederhana, yaitu:

$$S_{ij} = p - u = m \quad (17)$$

Persamaan (16) dan (17) mempunyai hasil yang sama, sekalipun diturunkan dari cara perhitungan yang berbeda. Kelebihan dari persamaan (17) adalah penurunan rumus yang digunakan lebih sederhana dan mudah.

Perbandingan aplikasi beberapa ukuran similaritas data biner

Validasi suatu ukuran similaritas salah satunya dapat dilakukan melalui perbandingan penerapan beberapa ukuran similaritas. Apabila penerapan beberapa ukuran similaritas memberikan hasil yang sama, dapat dikatakan bahwa ukuran similaritas tersebut valid. Pada tulisan ini akan diterapkan beberapa ukuran similaritas untuk kasus pengelompokan delapan rumah sakit negeri dan swasta (nama rumah sakit tidak disebutkan, melainkan digantikan dengan kode tertentu) berdasarkan ketersediaan fasilitas kamar kelas pertama. Data sekunder yang diambil pada tahun 2006

Tabel 2. Keberadaan Fasilitas Kamar Rumah Sakit di DIY untuk Kelas Satu

No	Kode Rumah Sakit	Fasilitas kamar kelas 1			
		Air panas	Televisi	AC	Layanan bel
1.	A	0	0	0	1
2.	B	0	1	1	0
3.	C	0	1	1	1
4.	D	1	1	1	1
5.	E	0	1	0	1
6.	F	1	1	1	0
7.	G	0	0	0	1
8.	H	0	1	1	1

untuk delapan rumah sakit di Daerah Istimewa Yogyakarta, adalah sebagaimana tertera pada Tabel 2. Dalam konteks permasalahan ini, kategori nol berarti tidak ada fasilitas, dan kategori satu berarti ada fasilitas.

Berdasarkan data Tabel 2., dengan menerapkan persamaan (2) dan (4), maka matriks simetris ukuran similaritas antara obyek *i* dengan obyek *j,* dinotasikan dengan matriks $(\mathbf{S}_{ij})$, untuk delapan Rumah Sakit adalah sebagai berikut:

(i) matriks ukuran similaritas Rogers – Tanimoto:

$$\mathbf{S}_{ij} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \\ G \\ H \end{array} \begin{array}{c} \begin{array}{cccccccc} A & B & C & D & E & F & G & H \end{array} \\ \begin{bmatrix} 1,00 & 0,14 & 0,33 & 0,14 & 0,60 & 0,00 & 1,00 & 0,33 \\ & 1,00 & 0,60 & 0,33 & 0,33 & 0,60 & 0,14 & 0,60 \\ & & 1,00 & 0,60 & 0,60 & 0,33 & 0,33 & 1,00 \\ & & & 1,00 & 0,33 & 0,60 & 0,14 & 0,60 \\ & & & & 1,00 & 0,14 & 0,60 & 0,60 \\ & & & & & 1,00 & 0,00 & 0,33 \\ & & & & & & 1,00 & 0,33 \\ & & & & & & & 1,00 \end{bmatrix} \end{array}$$

(ii) matriks ukuran similaritas Gower – Legendre:

$$\underline{S}_{ij} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \\ G \\ H \end{array} \begin{array}{c} \begin{array}{cccccccc} A & B & C & D & E & F & G & H \end{array} \\ \begin{bmatrix} 1,00 & 0,40 & 0,67 & 0,40 & 0,86 & 0,00 & 1,00 & 0,67 \\ & 1,00 & 0,86 & 0,67 & 0,67 & 0,86 & 0,40 & 0,86 \\ & & 1,00 & 0,86 & 0,86 & 0,67 & 0,67 & 1,00 \\ & & & 1,00 & 0,67 & 0,86 & 0,40 & 0,86 \\ & & & & 1,00 & 0,40 & 0,86 & 0,86 \\ & & & & & 1,00 & 0,00 & 0,67 \\ & & & & & & 1,00 & 0,67 \\ & & & & & & & 1,00 \end{bmatrix} \end{array}$$

Ukuran similaritas yang diusulkan Rogers – Tanimoto dan Gower – Legendre, mempunyai rentang nilai antara 0 dan 1, dimana semakin besar nilai similaritas, maka menunjukkan bahwa kedua obyek tersebut semakin mirip. Sebaliknya semakin kecil nilai similaritas, maka tingkat kemiripan kedua obyek semakin kecil. Ukuran similaritas yang diusulkan Rogers – Tanimoto dan Gower – Legendre, hanya berbeda pada bobot pembagi kategori dua obyek yang berbeda.

Dengan demikian kedua ukuran similaritas ini menghasilkan kesimpulan yang sama, yaitu rumah sakit A dan G, serta rumah sakit C dan H, mempunyai fasilitas yang sama untuk kamar kelas I. Sebaliknya, bahwa tidak ada satupun fasilitas kamar kelas I yang sama untuk rumah sakit A dengan F, serta rumah sakit F dengan rumah sakit G.

Sementara itu dengan menerapkan persamaan (16) dan (17) diperoleh matriks simetriks ukuran similaritas, dengan hasil yang sama, yaitu sebagai berikut:

(iii) matriks ukuran similaritas Widodo *et al*. dan matriks nilai *entropy*

$$\underline{S}_{ij} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \\ G \\ H \end{array} \begin{array}{c} \begin{array}{cccccccc} A & B & C & D & E & F & G & H \end{array} \\ \begin{bmatrix} 0 & 3 & 2 & 3 & 1 & 4 & 0 & 2 \\ & 0 & 1 & 2 & 2 & 1 & 3 & 1 \\ & & 0 & 1 & 1 & 2 & 2 & 0 \\ & & & 0 & 2 & 1 & 3 & 1 \\ & & & & 0 & 3 & 1 & 1 \\ & & & & & 0 & 4 & 2 \\ & & & & & & 0 & 2 \\ & & & & & & & 0 \end{bmatrix} \end{array}$$

Berkebalikan dengan makna pada ukuran similaritas Rogers – Tanimoto maupun Gower – Legendre, maka nilai *entropy* dan ukuran similaritas Widodo, dkk, mempunyai arti bahwa semakin besar nilainya menunjukkan

tingkat kemiripan yang semakin rendah. Sedangkan angka nol menunjukkan bahwa kedua obyek identik. Berdasarkan makna ini, maka kesimpulan yang sama dapat dikatakan bahwa rumah sakit A dan G, serta rumah sakit C dan H mempunyai fasilitas yang sama persis untuk kamar kelas I. Sementara itu rumah sakit A dan F, serta F dan G, mempunyai fasilitas yang sama sekali berbeda atau berkebalikan.

Sesuai dengan harapan perbandingan penerapan ukuran similaritas ini, maka untuk kasus ukuran similaritas antar rumah sakit di Yogyakarta, menghasilkan kesimpulan yang sama. Apabila dikembalikan pada konteks permasalahan semula, tentunya masyarakat yang akan menjalani rawat inap, dan menghendaki fasilitas di kamar kelas I, dengan alasan terdapat sejumlah rumah sakit (yaitu A dan G, atau C dan H) yang mempunyai fasilitas sama persis, masyarakat dapat memilih rumah sakit tersebut dengan mempertimbangkan harga sewa. Dalam konteks permasalahan ini harga kamar rawat inap setiap rumah sakit memang sengaja disembunyikan, dengan maksud ketika dua rumah sakit mempunyai fasilitas sama, tentunya harga kamar akan menjadi prioritas pertimbangan berikutnya. Secara prinsip, ukuran similaritas yang diturunkan dari usulan Widodo, dkk, ataupun diturunkan dari nilai *entropy,* dapat dijadikan sebagai alternatif validasi ukuran similaritas obyek bertipe biner.

KESIMPULAN

Peluang nilai kejadian dalam suatu kelompok dapat digunakan untuk mengukur similaritas antara obyek yang mempunyai bentuk data biner. Demikian halnya dengan cara yang sederhana, yaitu mempertimbangkan struktur pola kategori antara dua buah obyek, dapat digunakan dengan baik dan mudah untuk mengukur similaritas antara dua buah obyek dengan tipe data biner. Ukuran similaritas berbasiskan *entropy* ataupun berbasiskan pada bobot struktur pola kategori dua obyek, dapat dijadikan alternatif untuk validasi ukuran similaritas himpunan data biner.

DAFTAR PUSTAKA

Chen K & Liu L. 2005. *The "Bes K" for Entropy-based Categorical Data Clustering,* http://www.cc.gatech.edu/~kekechen/papers/catval05.pdf .

Deng S, Xu X & He Z. 2006. Improving Categorical Data Clustering Algorithm by Weighting Uncommon Attribute Value Matches*, ComSIS*, **3**(1)**:** 23 – 32.

Gower JC & Legendre P. 1986. Metric and Euclidean Properties of dissimilarity coefficient"s*, Journal of Classification*, **3**(1)**:** 5 – 48.

Hardle W & Simar L. 2007. *Applied Multivariate Analysis Statistical Analysis, Second Edition***,** Springer–Verlag.

Hair JF, Anderson RE, Tatham RL & Black, WG. 1995. *Multivariate Data Analysis with Reading* ***(***4^{nd} *ed)***,** New Jersey : Prentice-Hall.

Johnson RA & Wichern DW. 1992. *Applied Multivariate Statistical Analysis (*3^{nd} *ed)*, New Jersey : Prentice Hall.

Kim SY & Hamasaki T. 2008. Evaluation of Clustering on Preprocessing in Gene Expression Data**,** *International Journal of Biological*, 48 – 53.

Kudova P, Rezankova H, Huzek D & Snasel V. 2006. *Categorical Data Clustering Using Statistical Methods and Neural Networks***,** Proceedings of the Spring Young Researcher's Colloquium on Database and Information System, Moscow, Rusia.

Widodo E, Guritno S, Haryatmi S & Kariyam. 2009. *Ukuran Similaritas Data Kategorik Berbasiskan Pada Bobot Struktur Pola Kategori*, laporan penelitian internal Program Studi Statistika UII Yogyakarta.

Zorn C. 2003. *Agglomerative Clustering of Rangkings Data, with Application to Prison Rodeo Events,* Department of Political Sciense, Emory Universit, Atlanta, GA 30322, czorn@emory.edu.