MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **68** (2012) 581-592

ISSN 0340 - 6253

New Similarity Coefficients for Binary Data

Viviana Consonni* and Roberto Todeschini

Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences,
University of Milano-Bicocca, P.za della Scienza 1, I-20126 Milano, Italy
viviana.consonni@unimib.it, roberto.todeschini@unimib.it

(Received February 8, 2012)

Abstract

In the last few decades, the use of similarity measures has been becoming more and more important due to the relevance of comparing samples in order to find out clusters of similar samples, to generate priority lists, and, in general, to discover patterns in data structures.

In drug design, their relevance is already well established to search for the most suitable alternative to a target drug. In the QSAR field they are currently the key factor in read-accross strategy along with the defined chemical space.

Similarity indices for binary variables are usually called similarity coefficients and their first definitions date back to the end of the 19th century provided by scientists especially interested in taxonomic studies. Till date, more than 50 different similarity coefficients have been found in the literature, each having its own mathematical properties and characteristics and used in different scientific fields.

In this paper, five new similarity coefficients for binary data are proposed and compared with some well-known similarity coefficients.

1. Introduction

A great variety of data can be represented by binary variables [1;2], which express binary status of the sample, i.e. presence/absence, yes/no, true/false. For example, in archeology, binary data may denote that a particular artifact is found or not in a specific location; in taxonomy, binary data may denote the presence or absence of a particular taxonomic character in species; in psychology, binary data may denote if a person has a specific psychological trait; in chemistry, binary data may denote presence or absence in a molecule of a specific fragment or functional group.

Let two objects s and t be described by two binary vectors $\mathbf{x}$ and $\mathbf{y}$ each comprised of p variables with values 0/1. The binary similarity measures are commonly calculated from the data reported in Table 1, where a, b, c, and d are the frequencies of the events ($x = 1$ and $y = 1$), ($x = 1$ and $y = 0$), ($x = 0$ and $y = 1$), and ($x = 0$ and $y = 0$), respectively, in the pair of binary vectors describing the objects s and t; p is the total number of variables, equal to $a + b + c + d$, which is the length of each binary vector.

Table 1. Frequency table of the four possible combinations for two binary variables.

	$y = 1$		
$x = 1$	***a***	***b***	
$x = 0$	***c***	***d***	
		$b + d$	p

In other words,

- a is the number of variables equal to one for both objects (common "presences")
- d is the number of variables equal to zero for both objects (common "absences")
- $a + b$ is the number of variables equal to one for the *s-th* object
- $a + c$ is the number of variablese equal to one for the *t-th* object.

The diagonal entries a and d give information about the degree of similarity between the two objects, whereas the entries b and c give information about their dissimilarity.

There are two basic groups of similarity coefficients: symmetrical measures of similarity and asymmetrical measures of similarity. Symmetrical measures of similarity use both a and d, meaning that the double-zero state (d) for two objects (e.g. absence of a feature in both objects) is treated in exactly the same way as any other pair of values. These measures should be used when the double-zero state is a valid basis for comparing two objects; on the contrary, asymmetrical measures of similarity can be used, which skip the double-zero state in the similarity evaluation.

Some of the most common similarity coefficients are listed in Table 2.

2. New binary similarity coefficients

Five new similarity coefficients are proposed in this paper and are listed in Table 3.

As it can be seen from their formula, they are simply derived by applying the logarithm transformation to some common similarity coefficients and some their variants. All these coefficients range between 0 and 1.

The coefficients *T1*, *T3* and *T4* are obtained by the independent logarithm transformation of the numerator and denominator of Sokal-Michener (*SM*), Russel-Rao (*RR*) and Jaccard-Tanimoto (*Ja*) coefficients, respectively. *T2* is a logarithmic variant of *T1*, while *T5* is a logarithmic variant of the similarity coefficients based on correlation, i.e. ranging between -1 and +1. The denominator of *T5* corresponds to the maximum value the numerator can reach.

Table 2. Seven common binary coefficients.

Symbol	*Similarity coefficient*	*Name*	
Ja	$s_{Ja} = \frac{a}{a+b+c}$	Jaccard (1912) - Tanimoto	[3]
RR	$s_{RR} = \frac{a}{p}$	Russel – Rao (1940)	[4]
SM	$s_{SM} = \frac{a+d}{p}$	Sokal-Michener (1958), simple matching	[5]
SS1	$s_{SS1} = \frac{a}{a+2b+2c}$	Sokal – Sneath 1 (1963)	[6]
SS2	$s_{SS2} = \frac{2a+2d}{p+a+d}$	Sokal – Sneath 2 (1963)	[6]
SS3	$s_{SS3} = \frac{1}{4} \cdot \left[\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d} \right]$	Sokal – Sneath 3 (1963)	[6]
SS4	$s_{SS4} = \frac{a}{\sqrt{(a+b)(a+c)}} \cdot \frac{d}{\sqrt{(b+d)(c+d)}}$	Sokal – Sneath 4 (1963)	[6]

3. Data sets

In order to investigate the new similarity coefficients and compare them with existing ones, three simple simulated data sets (A, B, and C) have been generated, each comprised of a different number (20, 40, 60, respectively) of binary vectors of variable lenght: A (20x10), B (40x5), and C (60x10), respectively.

Along with the three simulated data sets, a real data set was taken from the literature. It consists of 125 pesticides, each described by 69 binary variables calculated by Dragon 6 software [7] and representing the presence/absence of atom pairs at different topological distances (between 1-4).

This data set was used to evaluate the performance of the considered similarity coefficients in structure similarity analysis of chemicals.

Table 3. The five new binary similarity coefficients.

Symbol	*New similarity coefficients*	*Derived from*
T1	$s_{T1} = \dfrac{log(1+a+d)}{log(1+p)}$	*SM*
T2	$s_{T2} = \dfrac{log(1+p)-log(1+b+c)}{log(1+p)}$	-
T3	$s_{T3} = \dfrac{log(1+a)}{log(1+p)}$	*RR*
T4	$s_{T4} = \dfrac{log(1+a)}{log(1+a+b+c)}$	*Ja*
T5	$s_{T5} = \dfrac{log(1+ad)-log(1+bc)}{log(1+p^2/4)}$	-

4. Comparison of similarity coefficients

Starting from the three simulated data sets, the similarity coefficients under investigation were calculated for each pair of objects, obtaining 190, 780 and 1770 similarity profiles for the data sets A, B, and C, respectively, Finally, all the similarity profiles were queued to generate a unique set constituted by 2740 records. Then, the matrix constituted by 2740 rows (pairs of objects) and 12 columns (the studied similarity coefficients) has been evaluated by Principal Component Analysis (PCA), correlation, and rank analyses.

In order to evaluate characteristics and relationships of the new similarity coefficients, a first comparison (Table 3) with seven of the most common coefficients (Table 2) was carried out by using PCA on the data set collecting different similarity values for 2740 pairs of objects.

The loading plots of the first four PCs, explaining 98.4% of the total variance, are shown in Fig. 1 and 2. These plots are useful to study the relationships among asymmetrical coefficients represented by triangles and symmetrical coefficients represented by circles; the new similarity coefficients are represented by squares.

The first component (PC1, Fig. 1) explains 78.5% of the total variance; all the loadings have almost the same values, meaning that this PC is related to the global degree of similarity between samples as measured by the different coefficients; this component is not particularly significant for the goal of this paper since it does not highlight differences among the similarity coefficients. The second component (PC2, Fig. 1) explains 14.1% of the variance and distinguishes symmetric (on the top) from asymmetric coefficients (on the bottom); indeed, positive loadings are related to symmetric coefficients such as simple matching (*SM*), second, third and fourth Sokal-Sneath coefficients (*SS2*, *SS3* and *SS4*), and the novel coefficients *T1*, *T2*, and *T5*, which are also symmetric with respect to *a* and *d* counts; negative loadings are related to asymmetric coefficients, i.e. first Sokal-Sneath (*SS1*), Jaccard-Tanimoto (*Ja*), and Russel-Rao (*RR*) coefficient, together with *T3* and *T4*. It is also noteworthy that the coefficients *T3* and *T4* are relatively isolated along PC2, thus revealing in some way a different kind of asymmetrical behaviour.

The third and fourth components (PC3 and PC4, Fig. 2) explain only 5.9% of the total variance and reveal some details about differences among the studied similarity coefficients. PC3 mainly explains some differences among symmetrical indices and specifically highlights the opposite behaviour of *SS2* and *T1* on the left side and *SS4* on the right side, whereas PC4 highlights the opposite behaviour of *SS1* and *T2* on the upper side and *T5* on the bottom side.

With the exception of *T1* and *SS2* coefficients, in the PC4 vs PC3 plot, all the other coefficients appear relatively isolated thus representing some specific information not explained by the other ones.

Relationships among the twelve similarity coefficients were further investigated by calculating the pairwise correlations of similarity coefficients by using both the Pearson formula (Table 4) and the Spearman rank correlation coefficients (Table 5).

From the Pearson correlations (Table 4), it results that the classical indices are largely correlated to each other with values always greater than 0.5. The same consideration holds for the new indices, with the exception for the correlation pair *T2* and *T3* (ρ = 0.370). Considering the relationships between old and new indices, the minimum correlation values are observed for the pairs *RR* and *T2* (ρ = 0.459), *SM* and *T3* (ρ = 0.475), and *SS2* and *T3* (ρ = 0.490). Note that a correlation background around 0.4-0.5 could be expected due to the common derivation of all the indices from only four parameters (*a*, *b*, *c*, and *d*).

Moreover, the maximum correlation between the new indices is observed for the pair *T3* and *T4* (ρ = 0.964), whereas the largest correlation between old and new similarity coefficients is found for the pair second Sokal-Sneath coefficient (*SS2*) and *T1* (ρ = 0.984). Also the correlations *SM-T1*, *SM-T2*, *RR-T3*, *Ja-T4*, *SS3-T5* are relatively high (ρ = 0.949, 0.949, 0.948, 0.948, and 0.968, respectively), as expected from the relationships implied by their definitions. Among the classical binary coefficients, *SM* and *SS2* also have high correlation equal to 0.983.

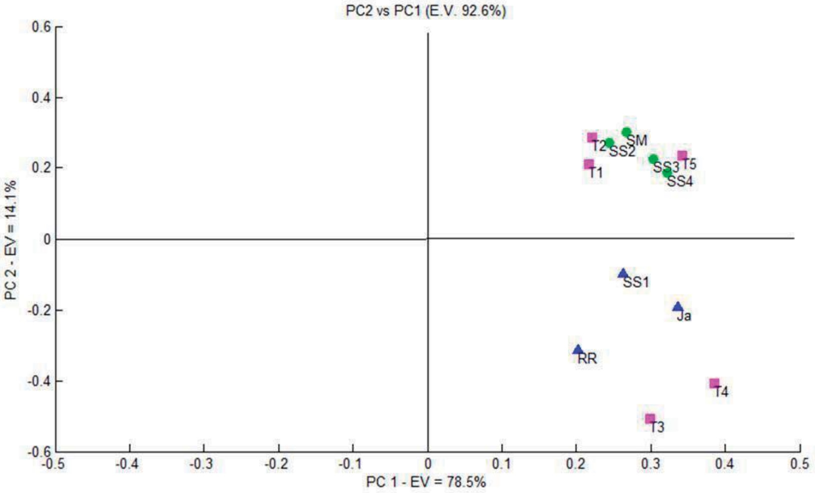


Fig. 1. Loading plot of PC2 vs PC1 of the similarity coefficients in analysis.

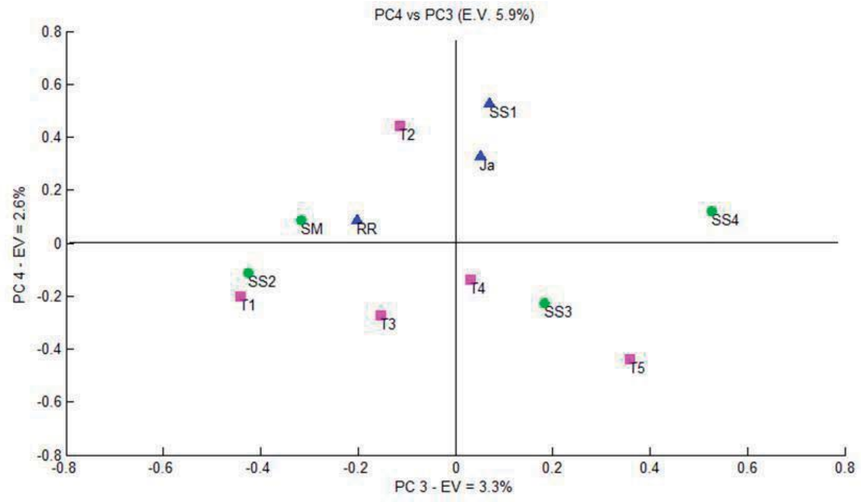


Fig. 2. Loadings of PC4 vs PC3 of the similarity coefficients in analysis.

Table 4. Pearson correlations between the 12 studied binary similarity coefficients. In boldface, the correlations greater than 0.98 and in italic the correlations greater than 0.90.

	SM	*Ja*	*RR*		*SS2*	*SS3*	*SS4*	*T1*	*T2*	*T3*	*T4*	*T5*
SM	1	0.745	0.519	0.752	**0.983**		0.832	*0.949*	*0.949*	0.475	0.620	0.854
Ja	0.745	1	0.893	*0.976*	0.706	0.808	0.832	0.688	0.730	0.872	*0.948*	0.777
RR	0.519	0.893	1	0.831	0.515	0.569	0.541	0.514	0.459	*0.948*	*0.903*	0.566
SS1		*0.976*	0.831	1	0.692	0.785	0.832	0.663	0.787	0.773	0.872	0.743
SS2	**0.983**	0.706	0.515	0.692	1	0.897	0.777	**0.984**	0.890	0.490	0.615	0.834
SS3	*0.912*	0.808	0.569	0.785	0.897	1	*0.944*		0.855	0.586	0.734	*0.968*
SS4		0.832	0.541	0.832	0.777	*0.944*	1	0.747	0.830		0.742	0.897
T1		0.688	0.514	0.663	**0.984**		0.747	1	0.827	0.523	0.634	0.795
T2	*0.949*	0.730	0.459	0.787	0.890	0.855	0.830	0.827	1	0.370	0.544	0.798
T3	0.475	*0.872*	*0.948*	*0.773*	0.490	0.586	0.569	0.523	0.370	1	*0.964*	0.581
T4	0.620	*0.948*	*0.903*	0.872	0.615	0.734	0.742	0.634	0.544	*0.964*	1	0.713
T5	0.854	0.777	0.566	0.743	0.834	*0.968*	0.897	0.795	0.798		0.713	1

Table 5. Spearman rank correlations between the 12 studied binary similarity coefficients. In boldface, the correlations greater than 0.98 and in italic the correlations greater than 0.90.

	SM	*Ja*	*RR*		*SS2*	*SS3*	*SS4*	*T1*	*T2*	*T3*	*T4*	*T5*
SM	1.000	0.737	0.516	0.737	**1.000**	0.881	0.811	**0.990**	**0.990**	0.494	0.701	0.858
Ja	0.737	1.000	*0.929*	**1.000**		0.796	0.801	0.747	0.712	*0.904*	**0.991**	0.781
RR		*0.929*	1.000	*0.929*		0.596	0.598	0.530	0.492	*0.987*	*0.946*	0.589
SS1	0.737	**1.000**	*0.929*	1.000	0.737	0.796	0.801	0.747	0.712	*0.904*	**0.991**	0.781
SS2	**1.000**		0.516	0.737	1.000	0.881	0.811		**0.990**	0.494	0.701	0.858
SS3		0.796	0.596	0.796	0.881	1.000	*0.965*		0.861	0.567	0.759	*0.976*
SS4	0.811	0.801	0.598		0.811	*0.965*	1.000	0.826	0.780	0.576	0.769	
T1	**0.990**	0.747	0.530	0.747	**0.990**	0.883	0.826	1.000	*0.961*	0.520	0.722	0.853
T2	**0.990**	0.712	0.492	0.712	**0.990**	0.861	0.780	*0.961*	1.000	0.458	0.666	0.846
T3	0.494	*0.904*	*0.987*	*0.904*	0.494	0.567	0.576	0.520	0.458	1.000	*0.938*	0.555
T4	0.701	**0.991**	*0.946*	**0.991**	0.701	0.759	0.769	0.722	0.666		1.000	0.739
T5	0.858	0.781	0.589		0.858	*0.976*	0.923	0.853	0.846	0.555	0.739	1.000

Similarity coefficients are frequently used to provide ranking of the objects (e.g. the most similar chemicals to a query compound). In this case, objects are ranked from the most to the less similar to the target and what is relevant is just the object ranking and not the strength of their similarity relationship. For this reason, the Spearman rank correlation analysis on the simulated data set was carried out replacing similarity values by ranks. Results of this analysis are reported in Table 5.

The rank correlations between the five novel coefficients are relatively low meaning that they produce quite different rankings of samples. Only the rankings provided by the coefficients *T1* and *T2* are quite similar ($\rho = 0.961$); moreover, both *T1* and *T2* have a high rank correlation equal to 0.990 with *SM* and *SS2*; the coefficients *T3* and *T4* have a rank correlation of 0.938 between them. Moreover, *T3* is highly correlated with *RR* ($\rho = 0.987$) and *T4* highly correlated with *Ja* and *SS1* ($\rho = 0.991$). The coefficient *T5* seems relatively correlated only to *SS3* and *SS4*, with a rank correlation of 0.976 and 0.923, respectively.

From the Table 5, it can be also noted that the ranks obtained by *SM* and *SS2* are identical, as well as the for the ranks obtained by *Ja* and *SS1*.

Finally, in order to investigate theperformance of the indices towards a practical application a structure similarity analysis was carried out on the real data set of 125 pesticides. Metobromuron was arbitrarily chosen as the reference molecule (Fig 3). Structure similarites of the remaining 124 chemicals towards metabromuron were then evaluated by the use of all the similarity indices in analysis.

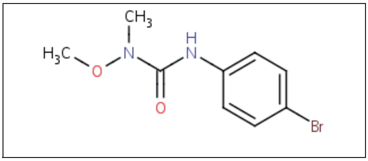


Fig. 3. Molecular structure of Metobromuron, selected as the reference compound for the similarity ranking of 124 pesticides.

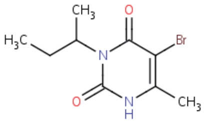
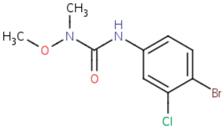
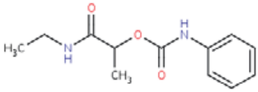
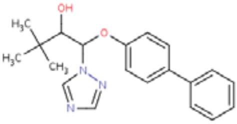
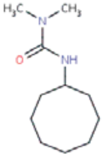
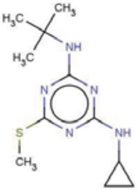
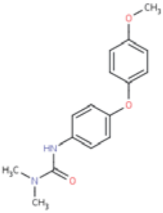
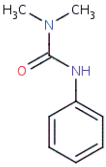
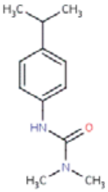
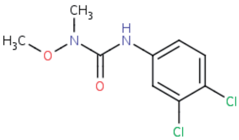
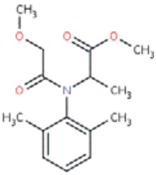
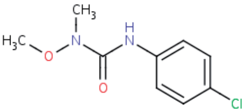
For each of the similarity coefficients, the ranking of pesticides from the most to the least similar to the query compound was derived. The list of the molecules ranked at the first 20 positions is reported in Table 6 for all the considered similarity coefficients. From the results of Table 6, it can be again concluded that indices *SM*, *SS2*, *T1*, and *T2* provide the same ranking of molecules (*SM* and *SS2* are theoretically correlated one); analogously, the same ranking is obtained by *Ja* and *SS1*, which are theoretically correlated one too, and by *RR* and *T3*, they being correlated one for this data set. An interesting conclusion can be drawn for indices *T4* and *T5* that seem to give rankings different from the other similarity coefficients; in particular, *T4* differs from *Ja* and *SS1* starting from the 9th rank position, while *T5* differs even from the first ranks.

The most similar molecules to Metobromuron as detected by the different similarity coefficients are shown in Table 7.

Table 6. First twenty rank positions of the 124 pesticide molecules for the twelve binary similarity coefficients.

Rank	*SM, SS2, T1, T2*	*Ja, SS1*	*RR, T3*	*SS3*	*SS4*	*T4*	*T5*
1	21, 29	21, 29	29	29	29	29	29
2			21	21	21	21	44
3	20, 25, 44, 71	20	69, 82	44	20	20	38, 53, 64
4		82		25, 71	25, 71	82	
5		25, 71	20, 120, 121			25, 71	
6				20	44		62
7	38, 53, 64, 82	44		38, 53, 64	82	44	76
8		69	9, 25, 28, 36, 45, 71, 72, 80, 99, 107		38, 53, 64	69	21
9						120, 121	19, 56, 86
10							
11	67			62	69	65, 66, 67, 72	
12				65, 66, 67	65, 66, 67		25, 71
13							
14							20
15	3, 52, 69, 72	72		69	62	38, 53, 64	3
16		120, 121		3	72		82
17				72	120, 121		65, 66, 67
18		62	18, 31, 33, 35, 39, 42, 43, 44, ...	52		9	
19	13, 33, 61, 85, 100, 110, 116, 120, 121	52		120, 121	52	33	
20		33			3	52	52

Table 7. Some of most similar pesticide molecules as found by the different similarity coefficients.

id 21 	*id 29* 	*id 25* 
id 20 		*id 62* 
id 44 	*id 53* 	*id 64* 
id 69 	*id 71* 	*id 82* 

5. Conclusions

In this paper five novel similarity coefficients were proposed. Simulated data were first generated to evaluate the novel indices in comparison with seven common indices.

From the Principal Component Analisys and the correlation analysis undertaken on the similarity data it was concluded that *T1* and *T2* do not provide new useful information since they behave as simple matching (*SM*) and second Sokal-Sneath (*SS2*) coefficients; *T3*, *T4* and *T5* are less correlated with the other common similarity indices, providing at least different similarity patterns, which are basically logarithmic transformations of some classical indices used for similarity analysis. Investigation of performances of the novel indices in structure similarity analysis of a real data set confirms that the same conclusions drawn from the analysis of simulated data and specifically that only indices *T4* and *T5* deserve further investigation to better understand their potentialto select alternative for a query compound.

Aknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under Grant Agreement n. 238701 of Marie Curie ITN Environmental Chemoinformatics (ECO) project.

References

[1] P. Legendre, L. Legendre, *Numerical Ecology*, Elsevier, Amsterdam, 1998, p. 854.

[2] V. Batagelj, M. Bren, Comparing resemblance measures, *J. Classif.* **12** (1995) 73-90.

[3] P. Jaccard, The distribution of the flora of the alpine zone, *New Phytologist* **11** (1912) 37-50.

[4] P. F. Russel, T. R. Rao, On habitat and association of species of *Anopheline larvae* in South Eastern Madras, *J. Malaria Inst. India* **3** (1940) 153-178.

[5] R. R. Sokal, C. D. Michener, A statistical method for evaluating systematic relationships, *Univ. Kansas. Sci. Bull.* **38** (1958) 1409-1438.

[6] R. R. Sokal, P. H. A. Sneath, *Principles of Numerical Taxonomy*, Freeman, San Francisco, 1963, p. 359.

[7] DRAGON (Software for Molecular Descriptor Calculation) -Version 6.0 - 2011 - Talete srl, http://www.talete.mi.it/