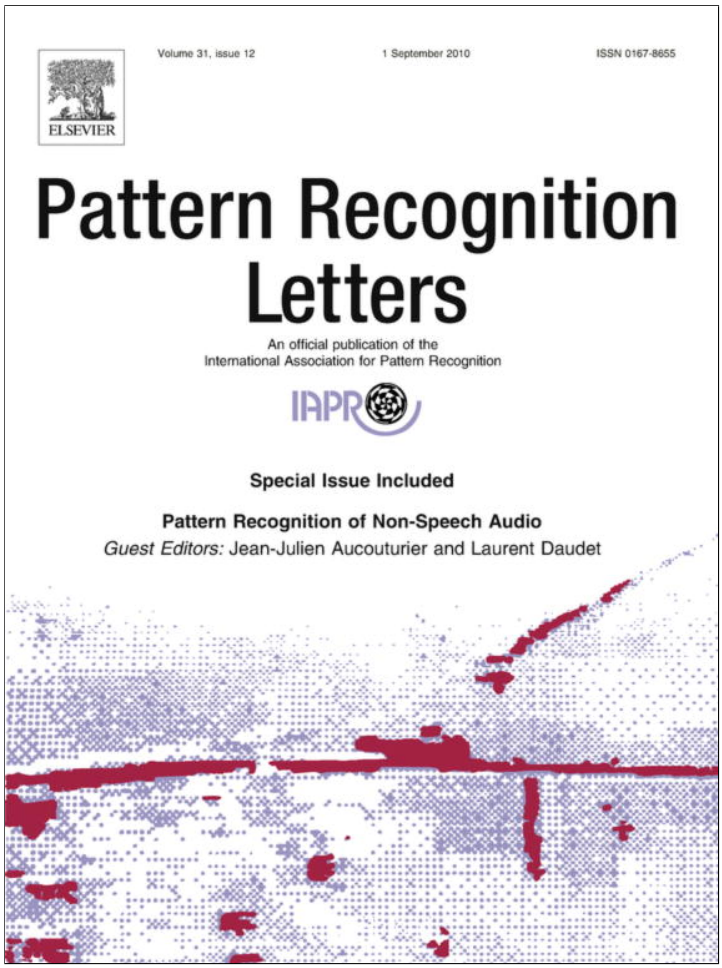

Pattern Recognition Letters 31 (2010) 1570–1576

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

The Positive Matching Index: A new similarity measure with optimal characteristics

Daniel Andrés Dos Santos [a,*], Reena Deutsch [b]

[a] CONICET – Facultad de Ciencias Naturales e Instituto Miguel Lillo, Universidad Nacional de Tucumán, Argentina
[b] Department of Psychiatry, University of California San Diego, San Diego, CA, USA

ARTICLE INFO

Article history:
Received 5 October 2009
Available online 18 March 2010
Communicated by W. Pedrycz

Keywords:
Binary data
Association coefficient
Jaccard index
Dice index
Similarity

ABSTRACT

Despite the many coefficients accounting for the resemblance between pairs of objects based on presence/absence data, no one measure shows optimal characteristics. In this work the Positive Matching Index (PMI) is proposed as a new measure of similarity between lists of attributes. PMI fulfills the Tulloss' theoretical prerequisites for similarity coefficients, is easy to calculate and has an intrinsic meaning expressable into a natural language. PMI is bounded between 0 and 1 and represents the mean proportion of positive matches relative to the size of attribute lists, ranging this cardinality continuously from the smaller list to the larger one. PMI behaves correctly where alternative indices either fail, or only approximate to the desirable properties for a similarity index. Empirical examples associated to biomedical research are provided to show outperformance of PMI in relation to standard indices such as Jaccard and Dice coefficients.

1. Introduction

A widespread task recognized in natural sciences is to compare pairs of items according to the lists of their attributes. As an example, we could refer to a common topic in the ecological literature, namely to compare the species composition of two areas. Similarity coefficients based on presence/absence data are frequently used to address this task. These measures evaluate the extent to which two items share categorical attributes codable into dummy variables, and they have been formulated in a number of slightly different ways and associated with even more authors: Jaccard, Czekanowski, Kulczinsky, etc. (Southwood and Henderson, 2000). There is no clear reason to prefer a given index over the others in all circumstances, and today the suggestion of Sneath and Sokal (1973) of subordinating the choice of an index to the research objectives and data is still appropriate.

On the grounds of theoretical prescriptions, simplicity and easy interpretation of results, there is no measure with the optimal characteristics. Tulloss (1997) proposed the Tripartite Similarity Index (TSI) to overcome drawbacks found on the pre-existing coefficients. For this, Tulloss (1997) used the guidelines of eight theoretical requirements highly conditioned by the work of Hayek (1994) and constructed a complicated formula oriented to fulfill them. The TSI is based on three cost functions and inherits concepts of manufacturing engineering. However, the TSI lacks a direct meaning and approaches to the theoretical properties in an approximate rather than an exact way. Moreover, Deutsch et al. (2006) found no beneficial performance of this measure in comparison to the simpler Dice coefficient to the extent of justifying its replacement based on multivariate binary responses in medical research. The aim of this paper is to present a new measure of similarity adjusted to all the theoretical requirements pointed out by Tulloss (1997), which is informative and easy to calculate.

2. The Positive Matching Index

The basic feature to obtain the association between objects based on presence-absence data is the following 2×2 contingency table:

		Item 2	
		Present	Absent
Item 1	Present	a	b
	Absent	c	d

The parameter a represents the number of common entries between lists (number of positive matches). Parameters b and c count the characteristics recorded for one of the two items under comparison. Finally, parameter d represents the number of descriptors that do not appear in either list. For the particular case of comparing sites based on species lists, negative matches d could be misleading

* Correspondence to: D.A.D. Santos, Facultad de Ciencias Naturales e Instituto Miguel Lillo, Universidad Nacional de Tucumán, Miguel Lillo 205, San Miguel de Tucumán, Tucumán, CP 4000, Argentina. Tel.: +54 381 4330633.
E-mail addresses: dadossantos@csnat.unt.edu.ar (D.A. Dos Santos), rdeutsch@ucsd.edu (R. Deutsch).

0167-8655/$ - see front matter © 2010 Published by Elsevier B.V.
doi:10.1016/j.patrec.2010.03.010

(Legendre and Legendre, 1998), since the absence of species at two sites may recognize very different underlying circumstances that preclude us to consider them as similar. The index we propose here discards the parameter d, and it will be called the Positive Matching Index (PMI) since its interest relies on the concordance of positive attributes.

Any assertion of non-zero similarity between objects requires that the objects do share characteristics. The proportion of common attributes yields an intuitive idea of similarity bounded on $\{0, 1\}$. Although there may be consensus on the numerator of this fraction, i.e. the cardinality of the set of common attributes, it is more difficult to establish its denominator in an uncontroversial way. Many similarity measures agree on this general remark, but they differentiate one another in the election of the denominator. Some measures ignore the asymmetry between sizes of the two input lists, choosing one of them as denominator of the ratio. Thus, the Simpson coefficient, $a/(a + \min(b, c))$, depends on the size of the smaller list, whereas the Braun–Blanquet coefficient, $a/(a + \max(b, c))$ divides common presences over the size of the larger list. The strategy of considering the smaller and larger lists together can be found on the Jaccard and Dice coefficients. The first one, $a/(a + b + c)$, calculates the ratio of common presences versus the total number of characteristics. The latter, $a/(a + 0.5(b + c)) = 2a/(2a + b + c)$, yields the proportion of positive matches for the mean size of the two lists.

The PMI also aims to retrieve the relationship between the parameter a and the sizes of the lists considered. When the two lists are equally sized, it is proposed that PMI should be supported by the following equation:

$$\mathrm{PMI}_{b=c} = \frac{a}{a+b} = \frac{a}{a+c} \tag{1}$$

However, when the lists are unequally sized ($b \neq c$), at least two possible proportions may be considered. The first one relates a with the size of the smaller list, whereas the second proportion is between a and the size of the larger list. It seems then reasonable to apply any averaging function to the different proportions. The Kulczinsky coefficient, $0.5(a/(a + b) + a/(a + c))$, moves along these guidelines in averaging both quotients referred above. But Kulczinsky coefficient has a problem: if all the entries of the smaller list appear in the larger, then the minimum value of the index is 0.5 despite the two lists being very disparate in size (Tulloss, 1997). Here, we propose to compute the average of the proportions of positive matches versus all the values contained between the cardinalities of the smaller and larger lists. Then, we suggest to gather the mean value of $f(x) = a/x$, x being a real number of the closed interval $[a + \min(b, c), a + \max(b, c)]$. Such a function is a rational one continuous on the specified interval, and the mean value can be obtained through the next formulae:

$$\begin{aligned}\mathrm{PMI}_{b\neq c} &= \frac{1}{(a+\max(b,c)) - (a+\min(b,c))} \int_{a+\min(b,c)}^{a+\max(b,c)} \frac{a}{x}\,dx \\ &= \frac{a}{|b-c|} \ln x\big|_{a+\min(b,c)}^{a+\max(b,c)} = \frac{a}{|b-c|} \ln\left(\frac{a+\max(b,c)}{a+\min(b,c)}\right)\end{aligned} \tag{2}$$

3. Adjustment of PMI to theoretical requirements

Tulloss (1997) offers a detailed framework to compare indices in an objective way. The author enumerates eight requirements and constructs the formula of the TSI in a way that strong deviations of those prescriptions are avoided. Furthermore, he explains how pre-existing measures do not meet some requirements. In this section, we will circumscribe to the behavior of the PMI and we reserve for the next section a synthetic and comparative overview of PMI in relation to some standard indices.

3.1. Requirement 1

A similarity index shall be sensitive to the relative size of the two lists to be compared; and great difference in size shall be interpreted to reduce the value of the similarity index (Tulloss, 1997). In other words, the coefficient should decrease as the difference $|b - c|$ becomes greater. The converse statement should not be predicated from this requirement as corollary, otherwise misleading measurements of similarity could be addressed. Thus, if two lists are balanced in size it does not follow that similarity should increase, because two lists may be completely disjoint despite being equally sized. Suppose we have two scenarios drawn from the triplet ($a > 0$, $b \geqslant 0$, $c > b$), where the difference between them relies on the value allocated for c, say $c_1 \ll c_2$. PMI goes down here because after taking derivatives, the decrease induced by c_2 at the left-hand factor of Eq. (2) proceeds at a higher rate than the respective rising promoted by the right-hand factor. In the limit, when $c_2 \to \infty$, the expression for PMI takes the indeterminate form of $\frac{\infty}{\infty}$. In applying L'Hôpital's Rule, PMI achieves its lower bound of zero.

3.2. Requirement 2

A similarity index shall be sensitive to the size of the sublist shared by a pair of lists; and an increase in difference in size between the smaller of the two lists and the sublist of common entries shall be interpreted to reduce the value of the similarity index (Tulloss, 1997). In order to assess if PMI is consistent with this requirement, we need to demonstrate that PMI decreases after adding any positive magnitude to the parameter of unique attributes of the smaller list. For example, suppose a is fixed and $b < c$, and that this relationship still holds after adding any positive quantity q to the parameter b (i.e. $0 < q < c - b$). Since PMI score corresponds here to the average value of a strictly decreasing function, namely the average fraction of shared attributes over the continuous interval bounded by the smallest and largest size of lists under comparison, any increasing in the cardinality of the smallest list induced by q implies that the values to be averaged are in the tail of the former curve, so PMI based on $b + q$ necessarily decreases and Requirement 2 is satisfied. Fig. 1 helps to demonstrate the

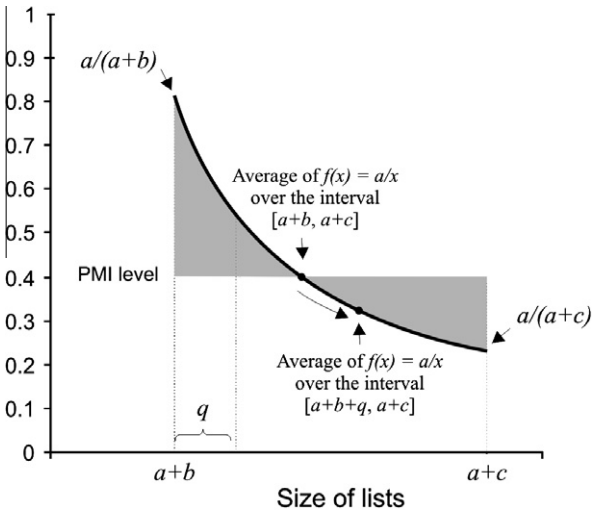

Fig. 1. Geometrical argument in favor of the adjustment of PMI to Tulloss' Requirement 2. PMI of the first triple data ($a > 0$, $b \geqslant 0$, $c > b$) is higher than the second triple data where q is added to b and $q < c - b$. Since PMI corresponds to the average value of $f(x) = a/x$, the gray areas above and below the PMI level are in balance. In adding q to the smaller list, PMI level necessarily should decrease in order to guarantee equivalence between areas.

inequality $\frac{a}{c-(b+q)}\ln\frac{a+c}{a+b+q} < \frac{a}{c-b}\ln\frac{a+c}{a+b}$, on the grounds of a geometrical reasoning. For the case of PMI based on b, its score is the height of the rectangle on the interval $a+b \leqslant x \leqslant a+c$ whose area equals the area under the curve $f(x) = a/x$. In other words, the area above PMI level inside the curve equals the area below PMI level outside the curve. In moving the left-hand extreme of the interval to a higher value $b+q$, both areas are in balance solely if PMI level decreases.

3.3. Requirement 3

A similarity index shall be sensitive to the percentage of entries in the larger list that are in common between the lists and to the percentage of entries in the smaller list that are in common between the two lists and shall increase as these two percentages increase (Tulloss, 1997). Here, we need to show that PMI varies positively in relation to greater proportions of positive matches in either list. That is, PMI increases whenever $a/(a+b)$ or $a/(a+c)$ increases. If $b = c$, this is true by definition. For $b < c$ and b, c fixed, when $a/(a+b)$ or $a/(a+c)$ increases, so does a. The derivative of the left-hand factor is positive, reflecting increasing values as a increases. A negative derivative of the right-hand factor reflects decreasing values. According to L'Hôpital's Rule, the limit of PMI as $a \to \infty$ is 1, hence the left-hand factor of PMI increases faster than the right-hand factor decreases, which returns a larger value of PMI as was to be proved since increasing a also results in an increase in $a/(a+b)$ and $a/(a+c)$. Similar results occur for $b > c$. Thus, as this index was developed in terms of the mean of these ratios along the continuous interval ranging from the smaller list to the larger one, any change applied to the proportions of common entries move in the expected direction.

3.4. Requirement 4

A similarity index shall yield values having fixed upper and lower bounds (Tulloss, 1997). It is easy to see that when lists are equally sized, $0 \leqslant \mathrm{PMI} \leqslant 1$ since a, b, and c are non-negative and the numerator is less than or equal to the denominator in all cases. In the situation of perfect matching ($a > 0$, $b = 0$, $c = 0$) or complete disassociation ($a = 0$, $b > 0$, $c > 0$), the upper and lower bounds are achieved, respectively. But lists can also be asymmetrical with the cardinalities of the smaller and larger lists divergent between them so that $|b - c|$ is non-zero and can tend to infinity. This can happen in each of the following cases: (1) ($a > 0$, $b > 0$, $c = 0$), (2) ($a > 0$, $b = 0$, $c > 0$), and (3) ($a = 0$, $b > 0$, $c > 0$). In cases (1) and (2), after applying L'Hôpital's Rule, as $a \to \infty$ and/or either $b \to \infty$ or $c \to \infty$, respectively, the limit of PMI approaches asymptotically to zero. In the third case, PMI equates to zero directly.

3.5. Requirement 5

A similarity index shall have the property that when two lists are identical, the similarity index for the two lists shall be equal to the upper bound of the index (Tulloss, 1997). As we already noted above, under this picture of complete matching ($a > 0$, $b = 0$, $c = 0$) PMI yields 1 since Eq. (1) becomes $a/(a+0)$.

3.6. Requirement 6

A similarity index shall have the property that when two lists have no entries in common, the similarity index for the lists shall be equal to the lower bound (Tulloss, 1997). Input data triples of the form ($a = 0$, $b > 0$, $c > 0$) are associated with this situation. Independently of being $b = c$ or $b \neq c$, PMI always yields the zero score, that is the lower bound.

3.7. Requirement 7

Distribution of values of a similarity index between zero and one shall be such that (a) if the size of two input lists is fixed, then the output shall vary roughly directly as the number of entries shared between the lists; and (b) if the smaller list is a subset of the larger list, then the value of the similarity index shall vary roughly inversely as the size of the larger list (Tulloss, 1997). The first part translates in that both the right-hand factor of Eq. (2) and the denominator of its left-hand factor remain constant, and say k is the product of these. So, PMI = ak and the direct relationship stated by item (a) is holding in an exact way. With regard to the item (b), suppose we face triples given as ($a > 0$, $b = 0$, $c > 0$), where a stays constant whereas c diverges from a. Here, we note that PMI reduces to the expression $\mathrm{PMI} = \frac{a}{c}\ln\left(1+\frac{c}{a}\right) = \ln\left(1+\frac{c}{a}\right)^{\frac{a}{c}}$. If we replace the quotient a/c by u, the last expression becomes $\mathrm{PMI} = \ln\left(1+\frac{1}{u}\right)^{u}$, that is the natural logarithm of the already known function giving rise to the Euler's number e as $u \to \infty$ ($c \to 0$, PMI = 1), and to the unity as $u \to 0$ ($c \to \infty$, PMI = 0) by using L'Hôpital's Rule. We are specially interested on the behavior of this function alongside the domain of c. PMI is here a strictly increasing function of u since the derivative of the previous function is strictly positive on the interval of interest $(0, \infty)$ (Strichartz, 2000). This implies that PMI assumes lower values as c gets larger ones, being supported thus the inverse relationship claimed for these magnitudes.

3.8. Requirement 8

A similarity index program shall check its input data to verify that the following relationships hold: $a + b > 0$ and $a + c > 0$ (Tulloss, 1997). This situation has been implicitly considered for the validity of Eqs. (1) and (2), otherwise the divide-by-zero problem would be present. According to Tulloss (1997) it makes no sense to perform a comparison between lists where one or both of them have no members. However, Deutsch et al. (2006) consider valuable that an index is able to deal with the trivial case where $a + b = 0$ and/or $a + c = 0$, at least in the medical setting. In order to circumvent the implementation caveat associated to Eqs. (1) and (2), the trivial case is directly treated as zero because PMI is concerned with the positive matches between lists. In the Appendix A we provide an *R* source code for calculating PMI with this last consideration in mind.

Tulloss (1997) provides some numerical examples given as triples (a, b, c), to show for TSI its (1) near invariance, (2) near linearity with regard to the variation in size of the set of shared entries, and (3) its inverse variation with regard to changes in size of the two lists. We adopt the several triples of Tulloss (1997) as a benchmark. Fig. 2 shows the responses of PMI exactly adjusted to the theoretical expectations in contrast to those of TSI.

4. Comparative performance of PMI and applications

4.1. A critical case

The TSI is an expression of three cost functions to mathematically address conflicting requirements of similarity (Deutsch et al., 2006). Penalties (reduction of the similarity metric) are applied when (1) there is more difference in the number of presences mutually exclusive and (2) there is an increase in the number of unique attributes of the smaller list. These two penalty cost functions are notated as U and S, respectively. On the contrary, a reward is obtained (similarity score rises up) when, for either item, there is an increasing proportion of positive matches relative to the pooled

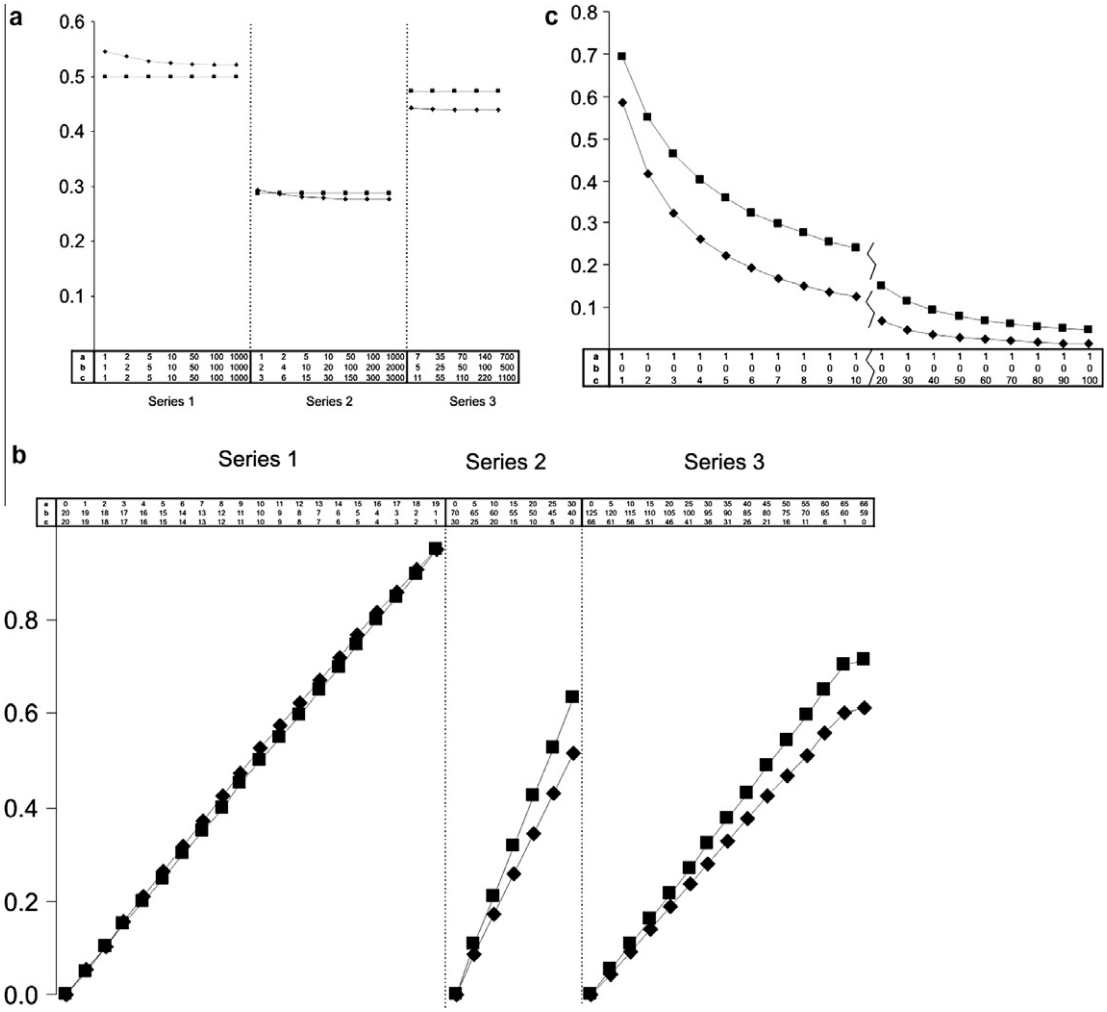

Fig. 2. Exact versus near adjustment to theoretical requirements of PMI and TSI, respectively. The properties considered are: (a) Invariance; (b) linear changes with regards to variations in the size of shared entries between lists; (c) inverse changes with regards to variation in size of the largest list. Values for input triple data (a, b, c) are in the x axis. Square (■) and diamond (♦) symbols correspond to responses of PMI and TSI, respectively.

total of positive records. The reward function is named R. Finally, TSI is the square root of the product $U \times S \times R$.

$$U = \log_2\left(1 + \frac{\min(b,c) + a}{\max(b,c) + a}\right), \quad 0 < U \leqslant 1$$
$$S = \frac{1}{\sqrt{\log_2(2 + \min(b,c)/(a+1))}}, \quad 0 < S \leqslant 1$$
$$R = \log_2(1 + \frac{a}{a+b}) \cdot \log_2\left(1 + \frac{a}{a+c}\right), \quad 0 < U \leqslant 1$$
$$\mathrm{TSI} = \sqrt{U \cdot S \cdot R} \tag{3}$$

Let us suppose we are comparing two assemblages of different scenarios represented by 501 triples (a, b, c), being a constant and $b \leqslant c$, responding to the following series: (500, 0, 1000), (500, 1, 999), ..., (500, 500, 500). The first scenario indicates to us that both assemblages share 500 species and that the smaller one is completely nested into the larger one, whereas the incidence lists of the last case reflects the maximal turnover possible given both a fixed count of shared species (500) and a joint total number of species (1500). Fig. 3 displays the values reported by the Jaccard, Dice, TSI and PMI indices under the different scenarios. The first two indices, as we have noted above, do not change along the series of triples since $b + c$ remains constant. In contrast, PMI and TSI do vary in accordance to changes on parameters b and c, but they follow opposite trends. The confusing behavior of TSI can be understood in terms of its inherent function U. TSI's U function correctly penalizes (reducing similarity scores) scenarios of very unbalanced sizes of input lists, but becomes a reward function rather than a neutral one when lists become equally sized. This role reversion of U function may dominate the similarity scoring and thus be misleading, since it disregards balanced scenarios with some trend to disassociation from balanced scenarios with high overlap of entries. This feature leads to TSI, for example, to judge the triple (5, 0, 100) less similar than (7, 93, 98) because of the more balanced condition of the second triple (not necessarily associated to higher similarity). Thus, TSI is outweighed by the function U.

4.2. Application 1: Neuropsychological example

Rippeth et al. (2004) studied the cognitive effects of concurrent HIV infection and methamphetamine dependence and found that both conditions may induce neuropsychological deficits. Rippeth et al. (2004) carried out a comprehensive, demographically corrected battery of tests to evaluate seven neurobehavioral domains: attention/working memory, verbal fluency, learning, recall, abstraction/problem solving, speed of information processing and motor skills. Study groups were comparable for age, education and ethnicity. Each test yielded a performance score susceptible

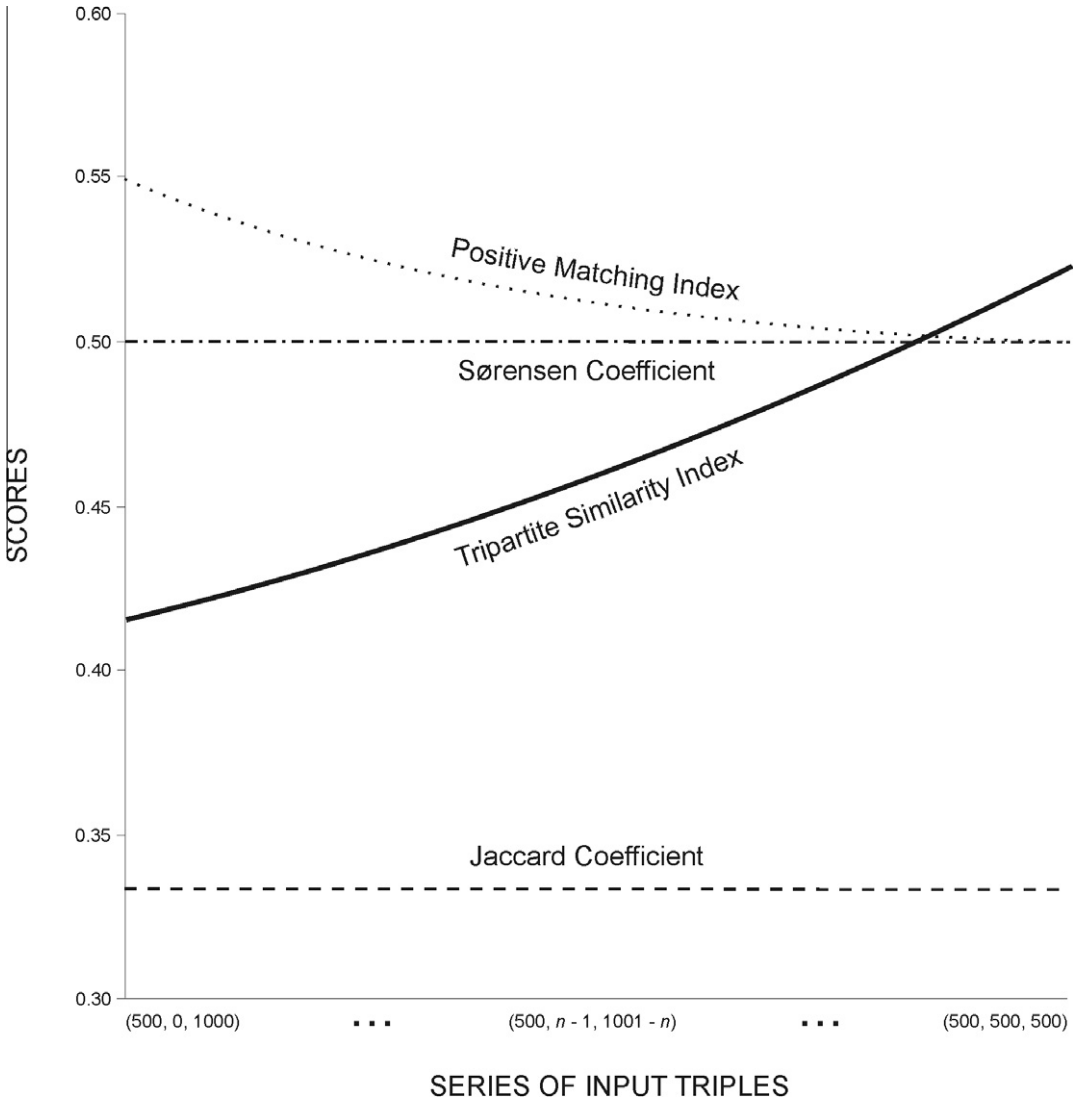

Fig. 3. Indices responses under different similarity scenarios, with fixed both joint total and number of shared entries. Scenarios are given as a series of 501 triples of parameters (a, b, c), starting with (500, 0, 1000) and ending at (500, 500, 500). Each new triple is obtained adding and subtracting the unity to b and c of the immediate preceding triple, respectively. In the x axis, n ranges from 1 to 501. Firstly, note the null sensitivity of the commonly used measures Jaccard and Dice. Lastly, note the opposite behavior of PMI and TSI.

of being converted later into a binary response after applying a given cutoff. This last procedure was used by Deutsch et al. (2006) in order to assess, in the cohort of HIV-negative subjects, if individuals classified as methamphetamine dependent (meth+) had similar neuropsychological performance but dissimilar to the pattern found outside this group, that is, individuals not dependent on methamphetamine (meth−). Deutsch et al. (2006) studied the multivariate binary profile of 68 meth+ and 47 meth− individuals for the seven neurological skills already mentioned. Dice, Jaccard, TSI and the simple matching coefficients were applied on data to obtain pairwise scores of similarity. Then, Deutsch et al. (2006) tested if the meth+ individuals were significantly more similar to each other than to individuals outside that group. For this, authors used the distinctness measure (Sokal and Rohlf, 1995) as the test statistic and performed a permutation test involving 1000 random arrangements of meth+ and meth− individuals into groups of 68 and 47 individuals as the original cluster sizes. Finally, they compared the observed distinctness value against the randomized distribution at a one-sided significance level. Table 1 summarizes the results of the permutation test, including now the PMI contribution

Table 1
Distinctness values for the cluster of subjects positive for the use of methamphetamine.

INDICES	Distinctness score	*P*-value
PMI	0.0818	0.0010
Dice	0.0722	0.0020
Jaccard	0.0531	0.0020
TSI	0.0038	0.4086

and ignoring the simple matching coefficient because this last measure is also sensitive to the parameter d (co-occurrence of absences), a feature explicitly disregarded by PMI and related measures under comparison. Despite minor numerical differences between our Table 1 and Deutsch et al. (2006, Table 3), attributable maybe to both random sampling noise and round off errors, they lead to the same conclusion: in opposition to the TSI, the simpler Dice and Jaccard coefficients enabled us to accept the hypothesis of similar pattern of cognitive functioning for the meth+ cluster. Noticeably, this conclusion still stays when PMI is implemented,

giving even more support to the hypothesis of a similar pattern of cognitive functioning.

4.3. Application 2: Myrmecological example

Longino et al. (2002) surveyed the ant fauna of La Selva Biological Station, Heredia Province, Costa Rica, through eight sampling methods falling into three categories: (1) canopy *Fogging* and *Malaise* traps sample the arboreal fauna; (2) *Berlese*, *Winkler*, *Barger*, and *Thompson* sample the soil and litter fauna; and (3) *Longino* and *Other* sample a combination of microhabitats. Given a local fauna like ants of La Selva Biological Station, it seems reasonable to expect a grouping of samples adjusted to these three categories delineated out by Longino et al. (2002), if we accept that different habitats produce a differential assemblage composition. The data set to be analyzed consists of 455 species $\times$ 8 samples incidence (presence–absence) matrix, and it is publicly available in ESA's Electronic Data Archive: *Ecological Archives* E083-011-A1 (http://www.esapubs.org/archive/ecol/E083/011/appendix-A.htm). From this data set, the similarity matrix $\boldsymbol{S} = [s_{ij}]$ is calculated, reflecting each entry s_{ij} the strength of association (measured with some similarity index) between ant sample methods i and j. As we are dealing with the indices Dice, Jaccard, TSI and PMI, we will obtain four similarity matrices. The next step was to evaluate, for each similarity matrix, the optimal flat partition of the eight methods into three blocks and compare these results with the expected partition {{*Fogging*, *Malaise*}, {*Berlese*, *Winkler*, *Barger*, *Thompson*}, {*Longino*, *Other*}}. The Stirling number of the second kind, i.e. the number of ways that n elements can be arranged into k non-empty subsets, for $n = 8$ sample methods and $k = 3$ blocks is 966, representing an accessible quantity to carry out an exhaustive search for the optimal partition. The similarities of each method to the others have been considered features or dimensions in the Euclidean space. The objective function to be minimized was the sum of squared distances of each item to the centroid of its respective membership block, a criterion compatible with the standard K-means method to produce flat partitions (e.g. Ball and Hall, 1967; MacQueen, 1967; Anderberg, 1973; Jain and Dubes, 1988). After enumerating all possible partitions with the R package **partitions** (Hankin and West, 2008), solely the matrix $\boldsymbol{S}$ based on PMI returned an optimal partition fully coincident with expectations (Fig. 4). The remaining similarity matrices, based on TSI, Jaccard and Dice coefficients, promoted the following partition as the optimal one: {{*Fogging*, *Malaise*, *Longino*, *Other*}, {*Berlese*, *Winkler*}, {*Barger*, *Thompson*}}. The important issue to remark here is the complete agreement between the expected partition of methods on the grounds of ecological foundations and that optimal partition induced by the PMI scores, a result not reached with alternative similarity measures.

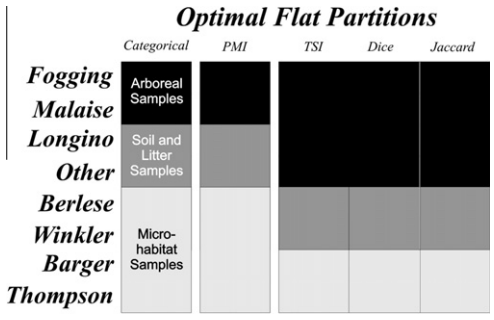

Fig. 4. Ants sample methods grouped into classes of similarity propinquity. The objective function for optimal partitions was the same as the K-means algorithm. Distances were calculated over the profile of cross similarities among methods. Complete enumeration of partitions of eight items into three blocks was carried out. Observe the perfect agreement between the partition of sample methods induced by the PMI similarity matrix and the categorical partition related to the type of explored habitat.

5. Discussion

The Jaccard index of similarity and the closely related Dice index are the two oldest and most widely used similarity indices for assessing compositional similarity of assemblages (Chao et al., 2005; Magurran, 2004). However, the selection of a particular index should not be based on subjective preferences or on previous widespread usage (Baselga et al., 2007). These indices are subject to aliasing, that is, they are prone to provide the same score under very different input data sets (Tulloss, 1997). The same index value can be obtained at very different pairs of parameters (b, c) if b and c yield the same value when summed, being thus the Requirement 1 not completely satisfied, because some insensitivity to the difference in size of the input data is exhibited for these measures.

A more sophisticated measure oriented to overcome shortcomings of alternative measures is the TSI. Although TSI may achieve numerical proxies to the values expected by the requirements, there is no rigorous proof of adjustment to all requirements. The critical case has suggested that TSI may judge as more similar, rather than less similar, a pair of lists equally sized but with some trend to disassociation of attributes, because of its associated U function rising up when lists become balanced in size to the extent of masking the decreasing trends of accompanying functions R and S.

PMI fulfills all the theoretical requirements of similarity measures raised by Tulloss (1997). The empirical examples have shown an acceptable performance of PMI, enabling to recover information considered *a priori* as intuitively reasonable. PMI is a measure easy to obtain and has an inherent meaning helping to the interpretation of results. Thus, if we are comparing two lists very unbalanced in size, say 10 and 100, and we obtain a PMI of 0.3, that result implies that both lists share the 30% of attributes on average along the domain of list sizes ranging from the smallest to the largest one. Future research direction is related to the extension of this measure to: (1) the analysis of ecological abundance tables, and (2) the search of similarities among geographical distributions in considering both the lists of areas where species occur in addition to the pattern of area occupancy.

Acknowledgements

This publication was supported by the National Council of Scientific and Technological Research of Argentina (CONICET). The authors wish to thank the HIV Neurobehavioral Research Center (P30 MH62512) at the University of California San Diego for supplying the dataset used in this article. E. Domínguez and M.C. Reynaga made useful suggestions in preparing the first draft of this manuscript.

Appendix A

The following R script (R Development Core Team, 2008) calculates the Positive Matching Index between the lists of features of two items. Parameters a (shared entries), b (unique positive entries of list 1) and c (unique positive entries of list 2) correspond to input variables of identical labels. The argument `dropReq8` is used to enable the comparison (TRUE) or not (FALSE) of lists without positive entries, that is, lists with all entries coded 0 throughout the set of attributes.

```
PMI <- function(a, b, c, dropReq8 = TRUE) {
  stopifnot(all(c(a, b, c) >= 0)) # Negative values are invalid.
  if (!dropReq8) stopifnot((a + b > 0) && (a + c > 0)) # Stop if both lists show no
positive entry
  if (a == 0) return(0) # No positive match occurs.
                        # Trivial case can be evaluated
                        # if Requirement 8 would be dropped out.
  if (b == c) return(a/(a + b)) # Equation (1).
  return(a/abs(b - c) * log((a + max(b, c))/(a + min(b, c)))) # Equation (2).
  }
```

References

Anderberg, M., 1973. Cluster Analysis for Applications. Academic, New York.

Ball, G., Hall, D., 1967. A clustering technique for summarizing multivariate data. Behav. Sci. 12, 153–155.

Baselga, A., Jiménez-Valverde, A., Niccolini, G., 2007. A multiple-site similarity measure independent of richness. Biol. Lett. 3, 642–645.

Chao, A., Chazdon, R.L., Colwell, R.K., Shen, T.-J., 2005. A new statistical approach for assessing similarity of species composition with incidence and abundance data. Ecol. Lett. 8, 148–159.

Deutsch, R., Cherner, M., Grant, I., 2006. Significance testing of a cluster of multivariate binary variables: Comparison of the tripartite T index to three common similarity measures. Statist. Meth. Med. Res. 15 (3), 285–299.

Hankin, R.K.S., West, L.J., 2008. Set partitions in *R*. J. Statist. Software 27.

Hayek, L.-A.C., 1994. Analysis of amphibian biodiversity data. In: Heyer, W.R., Donnelly, M.A., McDiarmid, R.W., Hayek, L.-A.C., Foster, M.S. (Eds.), Measuring and Monitoring Miological Diversity. Standard Methods for Amphibians. Smithsonian Institution, Washington, DC, pp. 207–269.

Jain, A., Dubes, R., 1988. Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, NJ.

Legendre, P., Legendre, L., 1998. Numerical Ecology, second ed. Elsevier Science BV, Amsterdam.

Longino, J.T., Coddington, J., Colwell, R.K., 2002. The ant fauna of a tropical rain forest: Estimating species richness three different ways. Ecology 83, 689–702.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: Proc. 5th Symposium on Mathematical Statistics and Probability, AD 669871, vol. 1, University of California Press, Berkeley, CA, pp. 281–297.

Magurran, A.E., 2004. Measuring Biological Diversity. Blackwell, Oxford.

Rippeth, J.D., Heaton, R.K., Carey, C.L., Marcotte, T.D., Moore, D.J., Gonzalez, R., Wolfson, T., Grant, I., The HNRC Group, 2004. Methamphetamine dependence increases risk of neuropsychological impairment in HIV infected persons. J. Intnat. Neuropsych. Soc. 10, 1–14.

R Development Core Team, 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Sneath, P.H.A., Sokal, R.R., 1973. Numerical Taxonomy. The Principles and Practice of Numerical Classification. W.H. Freeman, San Francisco.

Sokal, R.R., Rohlf, F.J., 1995. Biometry. The Principles and Practice of Statistics in Biological Research, third ed. W.H. Freeman and Co.. pp. 806–808.

Strichartz, R.S., 2000. The Way of Analysis, second ed. Jones and Bartlett Publishers. pp. 173–174.

Southwood, T.R.E., Henderson, P.A., 2000. Ecological Methods. Blackwell Malden, MA.

Tulloss, R.E., 1997. Assessment of similarity indices for undesirable properties and a new tripartite similarity index based on cost functions. In: Palm, M.E., Chapela, I.H. (Eds.), Mycology in Sustainable Development: Expanding Concepts Vanishing Borders. Parkway Publishers, pp. 122–143.