Int. J. Appl. Math. Comput. Sci., 2015, Vol. 25, No. 3, 645–657
DOI: 10.1515/amcs-2015-0047

amcs

ON THE ORDER EQUIVALENCE RELATION OF BINARY ASSOCIATION MEASURES

MARIUSZ PARADOWSKI [a]

[a] Department of Computational Intelligence
Wrocław University of Technology, Wyb. Wyspiańskiego 27, 50–370 Wrocław, Poland
e-mail: mariusz.paradowski@pwr.edu.pl

Over a century of research has resulted in a set of more than a hundred binary association measures. Many of them share similar properties. An overview of binary association measures is presented, focused on their order equivalences. Association measures are grouped according to their relations. Transformations between these measures are shown, both formally and visually. A generalization coefficient is proposed, based on joint probability and marginal probabilities. Combining association measures is one of recent trends in computer science. Measures are combined in linear and non-linear discrimination models, automated feature selection or construction. Knowledge about their relations is particularly important to avoid problems of meaningless results, zeroed generalized variances, the curse of dimensionality, or simply to save time.

Keywords: association coefficient, result ranking, linear combination, zeroed variance determinant, feature selection.

1. Introduction

Binary association measures, also known as association coefficients, have more than a hundred years of history of modern science. The Jaccard coefficient (Jaccard, 1912) may be listed among the oldest ones. Association measures have various origins, e.g., biology, taxonomy, psychology or language engineering. Their primary goal is to describe relations between pairs of objects sharing a common feature. During the last century, numerous new association measures have appeared. Some of them are well recognized and considered classic, e.g., the Dice coefficient (Dice, 1945). Having different sources of origin and interpretation, they often share some common properties. Association measures have been frequently reviewed and analyzed (Cheetham and Hazel, 1969; Wolda, 1981; Batagelj and Bren, 1995; Clarke *et al.*, 2006; Nieddu and Rizzi, 2007). Despite over a century of research, new association coefficients constantly appear (e.g., Washtell and Markert, 2009; Consonni and Todeschini, 2012).

Recent research and overview papers tend to present long lists of available binary similarity measures. At least a hundred of various measures may be found. These papers are usually related to computer science (Choi *et al.*, 2010) or various sub-domains of machine learning (Pecina, 2005; 2008; 2010; Pecina and Schlesinger, 2006). The appearance of long lists of coefficients in these research domains is not accidental. It is a rather popular but not always valid approach to automatically select or combine subsets of features out of large sets of available ones.

Binary association measures are often considered the basic ones. They can be further extended to *n-gram* measures (Petrović *et al.*, 2010) describing direct relations between n objects simultaneously. Another related research topic is the detection of *indirect* associations (e.g., Kazienko, 2009). Indirect association takes place if two objects share common features with other objects, called *transitive* ones.

Basic concepts. Binary association measures are in a large majority defined using four basic values (e.g., Batagelj and Bren, 1995; Nieddu and Rizzi, 2007; Choi *et al.*, 2010). These values represent the frequency of measured objects. Let $f(xy)$ represent a number of objects sharing both features x and y. Let $f(x\overline{y})$ represent a number of objects sharing feature x, but missing feature y. Let $f(\overline{x}y)$ represent a number of objects sharing feature y, but missing feature x. Finally, let $f(\overline{xy})$ be the number of object not sharing x or y. A widely accepted naming

convention is the following:

$$\begin{aligned} a &= f(xy), \quad b = f(x\overline{y}), \\ c &= f(\overline{x}y), \quad d = f(\overline{xy}). \end{aligned} \tag{1}$$

The above four values are frequently given in the form of a *contingency table* (e.g., Consonni and Todeschini, 2012).

Use of association measures. Binary association measures are an effective tool of measurement if two features (x and y) coexist in a set of objects. Various association measures have different interpretations and can be used in different scenarios. Several overviews exist in the literature, including formal analysis and interpretation (e.g., Cheetham and Hazel, 1969; Tan *et al*., 2004). Given a problem to solve, prior selection of an appropriate association measure is not always possible. In such cases, detailed statistical analysis of achieved results is required, (e.g., Washtell and Markert, 2009; Consonni and Todeschini, 2012).

One of prominent applications of association measures in computer science can be found in natural language processing. Association measures are used to detect compound, bi-gram (further generalized into n-grams) terms called *multiple word expressions* (e.g., phrasal verbs, technical terms) or *proper names* within huge text corpora. The key idea is to order generated bi-grams according to some association measure value. Most interesting bi-grams are expected to be at the top of the ranking. The difficulty is that bi-gram components of various multiple word expressions or proper names are associated in very diversified ways. Selection of an appropriate association measure is a very laborious task. As a consequence, supervised machine learning methods may be used to automate the process. The approach is often referred to as *learning to rank*. An extensive literature exists on the topic, including the works of Kekäläinen (2005), Liu (2009) or Chapelle and Wu (2010). Sets of association measures become input features of supervised learning methods (Pecina, 2005; 2008; 2010; Pecina and Schlesinger, 2006), such as *linear logistic regression*, *linear discriminant analysis*, *support vector machines* or *neural networks*. The procedure allows choosing a single measure or to produce a linear or non-linear combination of measures which model relations in the most accurate way.

It is worth mentioning that the presented binary association measures can be further extended into n-gram association measures. This can be done in at least two different ways. The first one is a direct reformulation of binary measures into n-gram measures, e.g., the Jaccard index (Segond and Borgelt, 2011). The second one is the usage of generic, n-gram measures parametrized by a binary measure. Such generic measures combine binary measurements on partial features into a single n-gram value (e.g., Petrović *et al*., 2010).

Goal. The goal of this work is to show that a subset of these association measures may be simplified. The paper focuses on order equivalence relations between coefficients and presents them in detail. Lots of them are monotone or anti-monotone functions of each other. Some of them are simple linear combinations of others. Thus, instead of using numerous coefficients to describe associations, a few of them will be sufficient. Such knowledge is especially valuable when coefficients are applied in machine learning methods. Some typical errors in multivariate statistical machine learning may be avoided. Usage of typical routines of automatic feature selection may be simplified.

Motivation. To motivate the presented analysis from a statistical point of view, a citation of Wichern's (2007, p. 131) book on multivariate statistics is appropriate: *This common practice of creating new variables that are sums of the original variables and then including them in the data set has caused enough lost time that we emphasize the necessity of being alert to avoid these consequences.*

To motivate the presented analysis from a machine learning point of view, a reference to the *curse of dimensionality* (Friedman, 1997) should be made. In highly dimensional data, single dimensions become less significant and less informative. Thus, an unnecessary increase of data dimensionality can often do more harm than good.

Yet another reason has purely practical character. It has been observed many times that less experienced machine learning students and researchers re-implement and re-examine various similarity measures with similar or identical properties. This study may be of help to these researchers and save a lot of valuable time.

Related work. Studies on the association measure value or ranking equality may be found in the literature. Generalized coefficients have been proposed as well. Cheetham and Hazel (1969) analyzed similarity measure convergence depending on the values of their components. They point out several relations between coefficients. One of the first generalized association coefficients was proposed by Tversky (1977). The *ratio model* is defined as

$$S_T = \frac{a}{a + \alpha b + \beta c}, \quad \alpha, \beta > 0. \tag{2}$$

Hubalek (1982) presented a complete list of 43 coefficients and showed that they are related by various transformations (linear, squares, logarithmic or trigonometric). Gower and Legendre (1986) gave two generalized coefficients:

$$\begin{aligned} S_\theta &= \frac{a+d}{a+d+\theta(b+c)}, \\ T_\theta &= \frac{a}{a+\theta(b+c)}, \quad \theta > 0. \end{aligned} \tag{3}$$

A detailed study of relations between binary similarity measures may be found in the work of Batagelj and Bren (1995). Following earlier works, the authors identify three groups (named **S**, **T** and **Q**) of coefficients sharing equal rankings. A comprehensive overview and a further generalization of association measures are given by Nieddu and Rizzi (2007). Coefficients S_θ and T_θ are special cases of their S_{NR} coefficient:

$$S_{NR} = \frac{a + \alpha d}{a + \beta d + \gamma(b + c)}, \tag{4}$$

$$\alpha \in \{0, 1\}, \quad \beta \in \{0, 1\}, \quad \gamma > 0. \tag{5}$$

Rifqi *et al*. (2008) presented an interesting discussion on relations between 10 coefficients and formed three groups. Hoang *et al*. (2009) identified a total of five groups of coefficients with repeatable rankings, but they address the issue only very briefly. They do not differentiate monotone and anti-monotone coefficient transformations, which results in reversed ranking of measured objects. Choi *et al*. (2010) presented an analysis of association measures according to the similarity of achieved results on randomly generated datasets. Association measures are hierarchically grouped and presented in the form of a dendrogram.

Contribution. There are two contributions in the paper. The first one is a detailed analysis of order equivalence relations of association measures. Following the motivation, the information on linear combinations of association measures is also given. To the best of our knowledge, this is the most extensive and detailed study of the topic up to date.

We also propose a generalized coefficient, named Φ. This very simple measure has a statistical background. It combines three probabilities: two *marginal probabilities* and the *joint probability*. The key difference between the proposed generalized association measure and other generalizations is that the former focuses on modeling association rankings instead of association values. Using the parametrized coefficient, we are able to generate an association order equal and opposite to at least 20 well known similarity measures. Usage of a single coefficient may simplify and bring more order to machine learning research and development. The coefficient may be also used in machine learning approaches, especially in feature construction routines. Further relations of the proposed coefficient with others are still to be discovered.

2. Relations between association measures

In this section we show that many association measures are related to each other. Some of them lead to identical association rankings (they are monotone functions) while others to reversed rankings (they are anti-monotone functions). Some results are trivial, but worth showing for the purpose of completeness. For clarity and completeness of the overview, results shown in earlier works are also presented. The most important earlier works are studies done by Cheetham and Hazel (1969), Hubalek (1982) as well as Batagelj and Bren (1995). We follow group names given in the latest study. Groups not defined by these studies are named in a similar way.

2.1. Further symbol definitions. The following relations may be defined on top of four basic values (see Eqn. (1)):

$$a + b = f(x), \quad a + c = f(y), \quad a + b + c + d = n. \tag{6}$$

The above values can also be given a probabilistic interpretation. Let us assume that joint probability $p(xy)$ and marginal probabilities $p(x)$ and $p(y)$ are estimated using relative frequency. Then we may write the following equations for joint probabilities:

$$p(xy) = \frac{a}{n}, \quad p(x\overline{y}) = \frac{b}{n}, \tag{7}$$

$$p(\overline{x}y) = \frac{c}{n}, \quad p(\overline{xy}) = \frac{d}{n}, \tag{8}$$

and for marginal probabilities:

$$p(x) = \frac{a + b}{n}, \quad p(y) = \frac{a + c}{n}, \tag{9}$$

$$p(\overline{x}) = \frac{c + d}{n}, \quad p(\overline{y}) = \frac{b + d}{n}. \tag{10}$$

2.2. List of coefficients. To get a proper reference for further discussion, a list of association coefficients is presented. The discussed coefficients are presented in Table 1. The list is limited only to these coefficients, which have rankings equal with other ones or can be represented by the proposed Φ coefficient (see Section 3). Short names and group assignments are given together with each coefficient. The last column contains parameter values of the proposed coefficient for all measures which can be formulated using Φ.

A very extensive list of coefficients is given by Choi *et al*. (2010). Other lists of coefficients can be found, among others, in the works of Cheetham and Hazel (1969), Hubalek (1982) as well as Pecina (2010).

2.3. Group RR. This group consists of the following coefficients: *Russel–Rao* (RR), *joint probability*, *generalized Nieddu* $S_{NR}(0, 1, 1)$ and *Consonni T3* (CT_3). They all generate an identical order of associated objects. However, Consonni T3 generates an identical association order if n is constant for all measurements.

Table 1. Definitions of the analyzed coefficients. The original article or appropriate coefficient survey is given for reference. Group assignment is presented (see coefficient transformations in the further part of this section). The superscript n represents group assignment if all measurements have equal $n = a + b + c + d$. The last column of the table shows the generalization using the proposed coefficient Φ (see Section 3).

Association measure name	Short	Definition	Group	Φ generalization
Joint Probability	JP	$p(xy)$	RR	$\Phi(1,0,0)$
Russel–Rao (Hubalek, 1982)	RR	$\frac{a}{a+b+c+d}$	RR	$\Phi(1,0,0)$
Consonni T3 (Consonni and Todeschini, 2012)	CT_3	$\frac{\log(1+a)}{\log(1+a+b+c+d)}$	RR^n	$\Phi(1,0,0)$
Nieddu (Nieddu and Rizzi, 2007)	S_{NR}	$\frac{a+\alpha d}{a+\beta d+\gamma(b+c)} : \alpha=0, \beta=1, \gamma=1$	RR	$\Phi(1,0,0)$
Sokal–Michiner (Hubalek, 1982)	SM	$\frac{a+d}{a+b+c+d}$	S	–
Hamann (Cheetham and Hazel, 1969)	Ham	$\frac{a+d-b-c}{a+b+c+d}$	S	–
Rogers–Tanimoto (Cheetham and Hazel, 1969)	RT	$\frac{a+d}{a+2b+2c+d}$	S	–
1st Sokal–Sneath (Hubalek, 1982)	SS_1	$\frac{a+d}{a+\frac{1}{2}b+\frac{1}{2}c+d}$	S	–
3rd Sokal–Sneath (Hubalek, 1982)	SS_3	$\frac{b+c}{a+d}$	S	–
Consonni T1 (Consonni and Todeschini, 2012)	CT_1	$\frac{\log(1+a+d)}{\log(1+a+b+c+d)}$	S^n	–
Consonni T2 (Consonni and Todeschini, 2012)	CT_2	$\frac{\log(1+a+b+c+d)-\log(1+b+c)}{\log(1+a+b+c+d)}$	S^n	–
Gower S_θ (Gower and Legendre, 1986)	S_θ	$\frac{a+d}{a+d+\theta(b+c)}, \theta>0$	S	–
Nieddu (Nieddu and Rizzi, 2007)	S_{NR}	$\frac{a+\alpha d}{a+\beta d+\gamma(b+c)} : \alpha=\beta=1, \gamma>0$	S	–
Jaccard (Jaccard, 1912)	J	$\frac{a}{a+b+c}$	T	$\Phi(1,0,-1)$
Dice (Dice, 1945)	D	$\frac{2a}{2a+b+c}$	T	$\Phi(1,0,-1)$
1st Kulczynski (Cheetham and Hazel, 1969)	K_1	$\frac{a}{b+c}$	T	$\Phi(1,0,-1)$
2nd Sokal–Sneath (Hubalek, 1982)	SS_2	$\frac{a}{a+2b+2c}$	T	$\Phi(1,0,-1)$
Anderberg (Duarte *et al*., 1999)	And	$\frac{a}{a+2b+2c}$	T	$\Phi(1,0,-1)$
Bray–Curtis (Clarke *et al*., 2006)	BC	$\frac{b+c}{2a+b+c}$	T	$\Phi(-1,0,1)$
Normalized expectation	NE	$\frac{2f(xy)}{f(x)+f(y)}$	T	$\Phi(1,0,-1)$
Tversky (Tversky, 1977)	S_T	$\frac{a}{a+\alpha b+\beta c} : \alpha=\beta>0$	T	$\Phi(1,0,-1)$
Gower T_θ (Gower and Legendre, 1986)	T_θ	$\frac{a}{a+\theta(b+c)}, \theta>0$	T	$\Phi(1,0,-1)$
Nieddu (Nieddu and Rizzi, 2007)	S_{NR}	$\frac{a+\alpha d}{a+\beta d+\gamma(b+c)} : \alpha=\beta=0, \gamma>0$	T	$\Phi(1,0,-1)$
Odds ratio	OR	$\frac{ad}{bc}$	Q	–
Yulle's $\mathcal{Q}$ (Cheetham and Hazel, 1969)	YQ	$\frac{ad-bc}{ad+bc}$	Q	–
Yulle's ω (Hubalek, 1982)	$Y\omega$	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$	Q	–
Driver–Kroeber (Hubalek, 1982)	DK	$\frac{a}{\sqrt{(a+b)(a+c)}}$	DK	$\Phi\left(1,-\frac{1}{2},0\right)$
Ochiai (Hubalek, 1982)	Och	$\frac{a}{\sqrt{(a+b)(a+c)}}$	DK	$\Phi\left(1,-\frac{1}{2},0\right)$
Otsuka (Cheetham and Hazel, 1969)	Ots	$\frac{a}{\sqrt{(a+b)(a+c)}}$	DK	$\Phi\left(1,-\frac{1}{2},0\right)$
Sorgenfrei (Hubalek, 1982)	$Sorg$	$\frac{a^2}{(a+b)(a+c)}$	DK	$\Phi(2,-1,0)$
Mutual dependency	MD	$\log\frac{p(xy)^2}{p(x)p(y)}$	DK	$\Phi(2,-1,0)$
Forbes (Hubalek, 1982)	F	$\frac{na}{(a+b)(a+c)}$	F	$\Phi(1,-1,0)$
Pairwise mutual information	PMI	$\log\frac{p(xy)}{p(x)p(y)}$	F	$\Phi(1,-1,0)$
Gilbert–Wells (Hubalek, 1982)	GW	$\log a - \log n - \log\frac{a+b}{n} - \log\frac{a+c}{n}$	F	$\Phi(1,-1,0)$
Confidence	C	$\max(p(y\|x), p(x\|y))$	C	–
Simpson (Cheetham and Hazel, 1969)	$Simp$	$\frac{a}{\min(a+b,b+c)}$	C	–
Phi (Cheetham and Hazel, 1969)	Phi	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	P	–
Pearson (Cheetham and Hazel, 1969)	$Pear$	$\frac{p(xy)-p(x)p(y)}{\sqrt{p(x)p(y)(1-p(x))(1-p(y))}}$	P	–
Log freq. biased MD (Pecina, 2010)	$fbMD$	$\log\frac{p(xy)^2}{p(x)p(y)} + \log p(xy)$	L	$\Phi(3,-1,0)$
FSCP (Buczyński, 2004)	$FSCP$	$\frac{a^3}{(a+b)(a+c)}$	L	$\Phi(3,-1,0)$
2nd Kulczynski (Cheetham and Hazel, 1969)	K_2	$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$	–	$\Phi(1,-1,1)$
Mutual expectation	ME	$\frac{2f(xy)}{f(x)+f(y)}p(xy)$	–	$\Phi(2,0,-1)$
Braun–Blanquet	BB	$\frac{1}{2}\frac{a}{max(a+b,a+c)}$	–	–

Statement 1. *Let $p(xy)$ be estimated using relative frequency. Then the Russel–Rao coefficient is equal to the joint probability $p(xy)$. It is also equal to the coefficient Nieddu $S_{NR}(0,1,1)$ (see Eqn. (4)).*

Proof. It is trivial, as we have

$$RR = \frac{a}{a+b+c+d} = S_{NR}(0,1,1) = \frac{a}{n} = p(xy).$$

■

Statement 2. *The coefficient Consonni T3 generates the association ranking equal to the Russel–Rao coefficient. The rankings are equal if n is constant for all observations. We show that, if*

$$CT_3(a_1,b_1,c_1,d_1) > CT_3(a_2,b_2,c_2,d_2), \tag{11}$$

then we have

$$RR(a_1,b_1,c_1,d_1) > RR(a_2,b_2,c_2,d_2). \tag{12}$$

Proof. Let $n = a_1+b_1+c_1+d_1 = a_2+b_2+c_2+d_2$. Multiply both the sides of Eqn. (11) by $\log(1+n)$:

$$\log(1+a_1) > \log(1+a_2). \tag{13}$$

Apply e^x and subtract 1 from both the sides:

$$a_1 > a_2. \tag{14}$$

Finally, divide both the sides by n:

$$\frac{a_1}{n} > \frac{a_2}{n}. \tag{15}$$

Thus we have

$$RR(a_1,b_1,c_1,d_1) > RR(a_2,b_2,c_2,d_2). \tag{16}$$

The corresponding transformations can be shown for $CT_3(a_1,b_1,c_1,d_1) = CT_3(a_2,b_2,c_2,d_2)$ and $CT_3(a_1,b_1,c_1,d_1) < CT_3(a_2,b_2,c_2,d_2)$. ■

2.4. Group S. This group consists of the following coefficients: *Sokal–Michiner* (SM), *Hamann* (Ham), *Rogers–Tanimoto* (RT), *first Sokal–Sneath* (SS_1), *third Sokal–Sneath* (SS_3), *Consonni T1* (CT_1), *Consonni T2* (CT_2), *generalized Gower* S_θ and *generalized Nieddu* S_{NR}. They all generate an identical order of the associated objects, except for the third Sokal–Sneath, which produces a reversed association order. However, Consonni coefficients generate an identical association order if and only if n is constant for all measurements.

The original group S (Batagelj and Bren, 1995) contained fewer coefficients. The following coefficients are added after a literature study: first Sokal–Sneath (see Hubalek, 1982), Consonni T1 and Consonni T2.

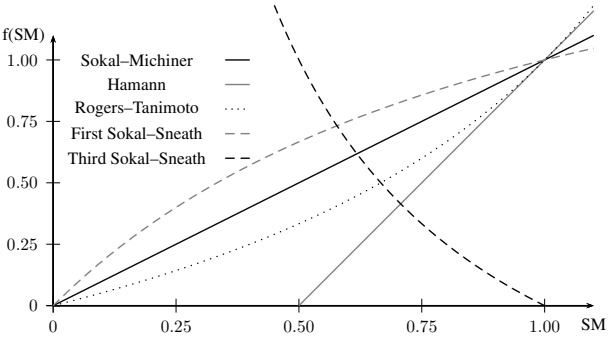

Fig. 1. Relations between Sokal–Michiner and other coefficients of group S.

Statement 3. *The Hamann coefficient is an affine transformation of the Sokal–Michiner coefficient. The orders of the associated objects in both the coefficients are equal.*

Proof. We have

$$\begin{aligned} 2(SM)-1 &= 2\frac{a+d}{a+b+c+d} - 1 \\ &= \frac{2a+2d-n}{a+b+c+d} \\ &= \frac{a+d-b-c}{a+b+c+d} = Ham. \end{aligned} \tag{17}$$

■

Due to an affine transformation between the Hamann and Sokal–Michiner coefficients, a simultaneous usage of both the coefficients as machine learning features is questionable.

Statement 4. *The Rogers–Tanimoto coefficient is a monotone function of the Sokal–Michiner coefficient. The orders of the associated objects in both the coefficients are equal.*

Proof. We have

$$\begin{aligned} \frac{2}{2-SM} - 1 &= \frac{2}{2-\frac{a+d}{a+b+c+d}} - 1 \\ &= \frac{2}{\frac{a+2b+2c+d}{a+b+c+d}} - 1 \\ &= \frac{a+d}{a+2b+2c+d} = RT. \end{aligned} \tag{18}$$

■

Statement 5. *The first Sokal–Sneath coefficient is a monotone function of the Sokal–Michiner coefficient. The orders of the associated objects in both the coefficients are equal.*

Proof. We have

$$\begin{aligned}\frac{-2}{1+SM}+2 &= \frac{-2}{1+\frac{a+d}{a+b+c+d}}+2\\ &= \frac{-2}{\frac{2a+b+c+2d}{a+b+c+d}}+2\\ &= \frac{a+d}{a+\frac{1}{2}b+\frac{1}{2}c+d}=SS_1.\end{aligned} \tag{19}$$

■

Statement 6. *The third Sokal–Sneath coefficient is an anti-monotone function of the Sokal–Michiner coefficient. The orders of the associated objects in both the coefficients are opposite.*

Proof. We have

$$\begin{aligned}\frac{1}{SM}-1 &= \frac{1}{\frac{a+d}{a+b+c+d}}-1\\ &= \frac{a+b+c+d}{a+d}-\frac{a+d}{a+d}\\ &= \frac{b+c}{a+d}=SS_3.\end{aligned} \tag{20}$$

■

Due to the anti-monotone relation between the third Sokal–Sneath and all the other coefficients of this group, simultaneous usage of these coefficients as machine learning features is questionable.

Statement 7. *The coefficient Consonni T1 generates the association ranking equal to the Sokal–Michiner coefficient. The rankings are equal if n is constant for all observations. We show that, if*

$$CT_1(a_1,b_1,c_1,d_1) > CT_1(a_2,b_2,c_2,d_2), \tag{21}$$

then we have

$$SM(a_1,b_1,c_1,d_1) > SM(a_2,b_2,c_2,d_2). \tag{22}$$

Proof. Let $n=a_1+b_1+c_1+d_1=a_2+b_2+c_2+d_2$. Multiply both the sides of Eqn. (21) by $\log(1+n)$:

$$\log(1+a_1+d_1) > \log(1+a_2+d_2). \tag{23}$$

Apply e^x and subtract 1 from the both sides:

$$a_1+d_1 > a_2+d_2. \tag{24}$$

Finally, divide both the sides by n:

$$\frac{a_1+d_1}{n} > \frac{a_2+d_2}{n}. \tag{25}$$

Thus we have

$$SM(a_1,b_1,c_1,d_1) > SM(a_2,b_2,c_2,d_2). \tag{26}$$

The corresponding transformations can be shown for $CT_1(a_1,b_1,c_1,d_1) = CT_1(a_2,b_2,c_2,d_2)$ and $CT_1(a_1,b_1,c_1,d_1) < CT_1(a_2,b_2,c_2,d_2)$. ■

Statement 8. *The coefficient Consonni T2 generates the association ranking equal to the Sokal–Michiner coefficient. The rankings are equal if n is constant for all observations. We show that, if*

$$CT_2(a_1,b_1,c_1,d_1) > CT_2(a_2,b_2,c_2,d_2), \tag{27}$$

then we have

$$SM(a_1,b_1,c_1,d_1) > SM(a_2,b_2,c_2,d_2). \tag{28}$$

Proof. Let $n=a_1+b_1+c_1+d_1=a_2+b_2+c_2+d_2$. Multiply both the sides of Eqn. (27) by $-\log(1+n)$ and add $\log(1+n)$:

$$\log(1+b_1+c_1) < \log(1+b_2+c_2). \tag{29}$$

Given $a+b+c+d=n$, we have

$$\log(1+n-a_1-d_1) < \log(1+n-a_2-d_2). \tag{30}$$

Apply e^x and subtract $1+n$ from both the sides:

$$-a_1-d_1 < -a_2-d_2. \tag{31}$$

Finally, multiply the result by $-n$:

$$\frac{a_1+d_1}{n} > \frac{a_2+d_2}{n}. \tag{32}$$

Thus we have

$$SM(a_1,b_1,c_1,d_1) > SM(a_2,b_2,c_2,d_2). \tag{33}$$

The corresponding transformations can be shown for $CT_2(a_1,b_1,c_1,d_1) = CT_2(a_2,b_2,c_2,d_2)$ and $CT_2(a_1,b_1,c_1,d_1) < CT_2(a_2,b_2,c_2,d_2)$. ■

2.5. Group T. This group consists of the following coefficients: *Jaccard* (J), *Dice* (D), *normalized expectation* (NE), *Anderberg* (And), *Bray–Curtis* (BC), *First Kulczynski* (K_1), *second Sokal–Sneath* (SS_2), *generalized Tversky* S_T, *generalized Gower* T_θ and *generalized Nieddu* S_{NR}. They all generate identical rankings of associations, except for the Bray–Curtis coefficient. The presented group is extended compared with the original proposal (Batagelj and Bren, 1995). The first addition is the normalized expectation, equivalent to the Dice coefficient. The second one is the Anderberg coefficient, equivalent to the second Sokal–Sneath. The third addition is the Bray–Curtis coefficient, which is closely related to the Dice coefficient.

Statement 9. *Dice (normalized expectation) is a monotone function of the Jaccard coefficient. The orders of the associated objects in both the coefficients are equal. The Bray–Curtis is an affine transformation of the Dice coefficient. Dice and Bray–Curtis have opposite orders of the associated object, and thus it is also an anti-monotone function of the Jaccard coefficient.*

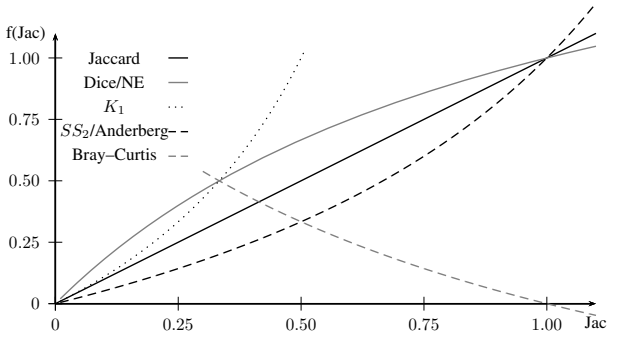

Fig. 2. Relations between Jaccard and other coefficients of group T.

Proof. We have

$$\begin{aligned} 2-\frac{2}{J+1} &= 2-\frac{2}{\frac{2a+b+c}{a+b+c}} \\ &= 2-\frac{2a+2b+2c}{2a+b+c} \\ &= \frac{2a}{2a+b+c} \\ &= \frac{2f(xy)}{f(x)+f(y)} = D = NE. \end{aligned} \tag{34}$$

Then

$$\frac{2}{J+1}-1 = \frac{2}{\frac{2a+b+c}{a+b+c}}-1 = \frac{b+c}{2a+b+c} = BC. \tag{35}$$

■

Statement 10. *The first Kulczynski coefficient is a monotone function of the Jaccard coefficient. The orders of the associated objects in both the coefficients are equal.*

Proof. We have

$$\begin{aligned} \frac{1}{1-J}-1 &= \frac{1}{1-\frac{a}{a+b+c}}-1 \\ &= \frac{1}{\frac{b+c}{a+b+c}}-1 \\ &= \frac{a+b+c}{b+c}-\frac{b+c}{b+c} = \frac{a}{b+c} = K_1. \end{aligned} \tag{36}$$

■

Statement 11. *The second Sokal–Sneath (Anderberg) coefficient is a monotone function of the Jaccard coefficient. The orders of the associated objects in both the coefficients are equal.*

Proof. We have

$$\begin{aligned} \frac{2}{2-J}-1 &= \frac{2}{2-\frac{a}{a+b+c}}-1 \\ &= \frac{2}{\frac{a+2b+2c}{a+b+c}}-1 \\ &= \frac{2a+2b+2c}{a+2b+2c}-1 \\ &= \frac{a}{a+2b+2c} = SS_2. \end{aligned} \tag{37}$$

■

2.6. Group Q. This group consists of the following coefficients: *odds ratio* (OR), *Yulle's* ω ($Y\omega$), *Yulle's* $\mathcal{Q}$ (YQ). They all generate identical orders of the associated objects. This group remains identical, as shown by Batagelj and Bren (1995).

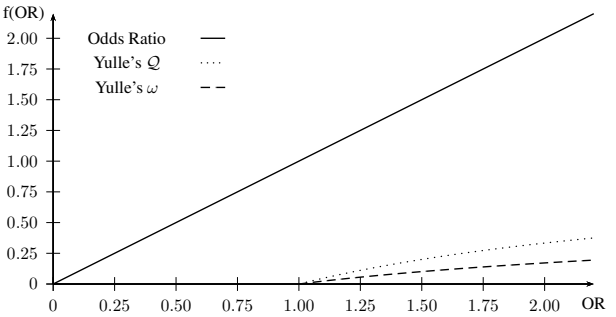

Fig. 3. Relations between the odds ratio and other coefficients of group Q.

Statement 12. *Yulle's* $\mathcal{Q}$ *coefficient is a monotone function of the odds ratio coefficient. The orders of the associated objects in both the coefficients are equal.*

Proof. We have

$$\begin{aligned} 1-\frac{2}{OR+1} &= 1-\frac{2}{\frac{ad}{bc}+1} \\ &= 1-\frac{2bc}{ad+bc} \\ &= \frac{ad-bc}{ad+bc} = YQ. \end{aligned} \tag{38}$$

■

Statement 13. *Yulle's* ω *coefficient is a monotone function of the odds ratio coefficient. The orders of the associated objects in both the coefficients are equal.*

Proof. We have

$$\begin{aligned} 1-\frac{2}{\sqrt{OR}+1} &= 1-\frac{2}{\frac{\sqrt{ad}}{\sqrt{bc}}+1} \\ &= 1-\frac{2\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}} \\ &= \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}} = Y\omega. \end{aligned} \tag{39}$$

■

2.7. Group DK. This group consists of the following coefficients: *Driver–Kroeber* (DK), *mutual dependency* (MD) and *Sorgenfrei* ($Sorg$). The Driver–Kroeber coefficient is also known as the *Ochiai* coefficient and the *Otsuka* coefficient. They all generate identical rankings of associations.

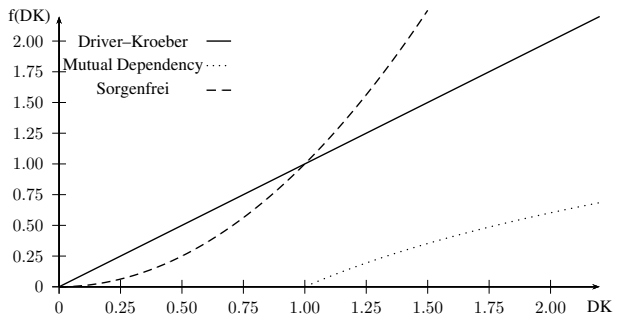

Fig. 4. Relations between Driver–Kroeber and other coefficients of group DK.

Statement 14. *Let $p(xy)$, $p(x)$ and $p(y)$ be estimated using relative frequency. Then the Driver–Kroeber coefficient is a monotone function of mutual dependency. The orders of the associated objects in both the coefficients are equal.*

Proof. We have

$$\begin{aligned} \log\left(DK^2\right) &= \log\left[\left(\frac{a}{\sqrt{(a+b)(a+c)}}\right)^2\right] \\ &= \log\left[\left(\frac{\frac{1}{n}a}{\sqrt{\frac{1}{n^2}(a+b)(a+c)}}\right)^2\right] \\ &= \log\frac{p(xy)^2}{p(x)p(y)} = MD. \end{aligned} \tag{40}$$

■

Statement 15. *The Sorgenfrei coefficient is a monotone function of the Driver–Kroeber coefficient. Association rankings of both the coefficients are equal.*

Proof. It is trivial, as we have

$$DK^2 = \frac{a^2}{(a+b)(a+c)} = Sorg.$$

■

2.8. Group F. This group consists of the following coefficients: *pairwise mutual information* (PMI), *Forbes* (F) and *Gilbert–Wells* (GW). The group contains coefficients defined as a logarithm transformation of the Forbes coefficient. They all generate identical rankings of associations.

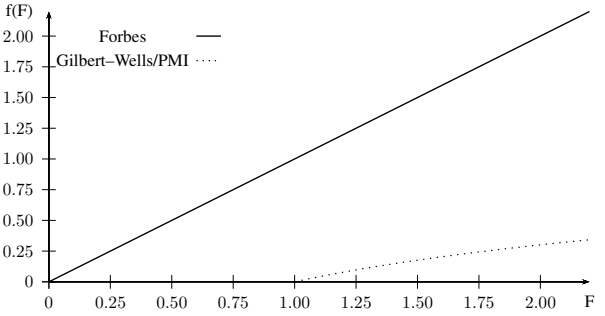

Fig. 5. Relations between Forbes and other association coefficients of group F.

Statement 16. (Hubalek, 1982) *Let $p(x)$, $p(y)$ and $p(xy)$ be estimated using relative frequency. Then the pairwise mutual information coefficient is equal to the Gilbert–Wells coefficient.*

Proof. We have

$$\begin{aligned} GW &= \log a - \log n - \log\frac{a+b}{n} - \log\frac{a+c}{n} \\ &= \log\frac{\frac{a}{n}}{\frac{(a+b)(a+c)}{n^2}} \\ &= \log\frac{p(xy)}{p(x)p(y)} = PMI. \end{aligned} \tag{41}$$

■

Statement 17. *Let $p(x)$, $p(y)$ and $p(xy)$ be estimated using relative frequency. Then the pairwise mutual information is a monotone function of the Forbes coefficient. Association rankings of both the coefficients are equal.*

Proof. We have

$$\begin{aligned} \log F &= \log\frac{na}{(a+b)(a+c)} \\ &= \log\frac{\frac{a}{n}}{\frac{a+b}{n}\frac{a+c}{n}} \\ &= \log\frac{p(xy)}{p(x)p(y)} = PMI. \end{aligned} \tag{42}$$

■

2.9. Other groups.

Statement 18. *Let $p(xy)$, $p(x)$ and $p(y)$ be estimated using relative frequency. Then the Simpson (Sim) coefficient is equal to the confidence (C) coefficient.*

Proof. We have

$$\begin{aligned} Sim &= \frac{a}{\min(a+b, b+c)} \\ &= \frac{p(xy)}{\min(p(x), p(y))} \\ &= \max\left(p(y|x), p(x|y)\right) = C. \end{aligned} \tag{43}$$

■

Statement 19. (Cheetham and Hazel, 1969) *Let $p(x)$, $p(y)$ and $p(xy)$ be estimated using relative frequency. Then the Pearson ($Pear$) coefficient is equal to the Phi (Phi) coefficient.*

Proof. We have

$$\begin{aligned} Pear &= \frac{p(xy) - p(x)p(y)}{\sqrt{p(x)p(y)(1-p(x))(1-p(y))}} \\ &= \frac{\frac{an}{n^2} - \frac{(a+b)(a+c)}{n^2}}{\sqrt{\frac{(a+b)(a+c)(n-a-b)(n-a-c)}{n^4}}} \\ &= \frac{a^2 + ab + ac + ad - a^2 - ab - ac - bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}} \\ &= \frac{ad - bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}} = Phi. \end{aligned} \tag{44}$$

■

2.10. Linear combinations. The last group of coefficients includes those that may be represented as linear combinations of others. Such knowledge is important when designing recognition methods. Some recognition methods (e.g., Fisher's linear discriminant, multivariate naive Bayes) are sensitive to linear combinations of features due to a problem with the zero generalized variance and inversion of the covariance matrix. Others (e.g., the simple perceptron, neural networks) are able to easily generate linear feature combinations. In such cases, linearly combined features may be simply discarded.

Statement 20. *Let $p(xy)$ and $p(\overline{xy})$ be estimated using relative frequency. Then the Sokal–Michiner coefficient is equal to the sum of joint probabilities.*

Proof. It is trivial, as we have

$$\begin{aligned} SM &= \frac{a+d}{a+b+c+d} \\ &= \frac{a}{n} + \frac{d}{n} = p(xy) + p(\overline{xy}). \end{aligned} \tag{45}$$

■

Interesting questions may be raised here. Should group **S** be discarded and elementary probabilities $p(xy)$ and $p(\overline{xy})$ be used instead? In which machine learning methods is such an approach reasonable?

Statement 21. *Let $p(x)$, $p(y)$ and $p(xy)$ be estimated using relative frequency. Then the second Kulczynski coefficient is an average of conditional probabilities.*

Proof. It is trivial, as we have

$$\begin{aligned} K_2 &= \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right) \\ &= \frac{1}{2}\left(p(y|x) + p(x|y)\right). \end{aligned} \tag{46}$$

■

Statement 22. *The second Kulczynski coefficient is a linear combination of the Braun–Blanquet (BB) and Simpson (Sim) coefficients.*

Proof. We have

$$\begin{aligned} &\frac{1}{2}(BB + Sim) \\ &\quad= \frac{\frac{1}{2}a}{\max(a+b, a+c)} + \frac{\frac{1}{2}a}{\min(a+b, a+c)} \\ &\quad= \frac{a[\min(a+b, a+c) + \max(a+b, a+c)]}{2\max(a+b, a+c)\min(a+b, a+c)} \\ &\quad= \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right) = K_2. \end{aligned} \tag{47}$$

■

3. Generalized Φ coefficient

A generalized binary association coefficient named Φ is proposed. It can be used instead of several frequently employed ones. The main application area of the coefficient are machine learning methods, especially feature selection and feature generation ones. The proposed coefficient is defined using elementary probability values of events x and y. It combines the joint probability $p(xy)$, the marginal probabilities $p(x)$ and $p(y)$ and the mean of the marginal probabilities $\frac{1}{2}(p(x) + p(y))$.

The proposed generalized Φ coefficient is conceptually different from generalizations proposed by Tversky (1977), Gower and Legendre (1986) or Nieddu and Rizzi (2007). They focused on generalizing the definition based on a, b, c and d elementary object counts. The proposed coefficient is based on composite, probabilistic components. The Φ coefficient is defined as

$$\begin{aligned} &\Phi(\alpha, \beta, \gamma) \\ &\qquad= p(xy)^{\alpha} p(x)^{\beta} p(y)^{\beta} \left(\frac{p(x) + p(y)}{2}\right)^{\gamma}, \end{aligned} \tag{48}$$

where

$$\alpha, \beta, \gamma \in \mathbb{R}, \quad \Phi(\alpha, \beta, \gamma) \geq 0. \tag{49}$$

The proposed coefficient allows generating both normal and reversed orders of associations. It has the following property:

$$\Phi(\alpha, \beta, \gamma) = \frac{1}{\Phi(-\alpha, -\beta, -\gamma)}. \tag{50}$$

Given that $p(xy)$, $p(x)$ and $p(y)$ are estimated using relative frequency, $\Phi(\alpha, \beta, \gamma)$ is equal to

$$\begin{aligned}&\Phi(\alpha, \beta, \gamma)\\ &= \left(\frac{a}{n}\right)^{\alpha} \left(\frac{a+b}{n}\right)^{\beta} \left(\frac{a+c}{n}\right)^{\beta} \left(\frac{a+\frac{1}{2}(b+c)}{n}\right)^{\gamma}.\end{aligned} \tag{51}$$

3.1. Motivation and interpretation. Our goal is to define a coefficient that covers a large number of groups identified in Table 1. On the other hand, the coefficient should be as simple as possible and have the smallest possible number of parameters. After manual analysis of a number of available coefficients, three prominent components of typical association measures have been identified, namely, $p(xy)^{\alpha}$, $p(x)^{\beta}p(y)^{\beta}$ and $\frac{1}{2}\left(p(x)+p(y)\right)^{\gamma}$. The first one is typically used as a numerator, the second and third ones are typical denominators. Yet another expected property is the ability to generate both normal and reversed rankings of associations (see Eqn. (50)).

An interpretation of the generalized coefficient Φ comes together with specified parameter values. A parametrized Φ coefficient models a specific coefficient, and thus it may be interpreted as this coefficient. However, given the defined parameters, interpretation may still vary. For example, the Jaccard and Dice coefficients belong to the same group T (see Table 1) and are generated by the same set of parameters, i.e., $\Phi(1, 0, -1)$. Despite this fact, the Jaccard coefficient has the probability of the sum of events in the denominator, while the Dice coefficient has the sum of marginal probabilities. As a consequence, interpretations of both the coefficients is different. The key observation is that both the coefficients are monotone transformations of each other. Regardless of their interpretation, they still provide the same order of elements.

An interpretation of the coefficient Φ is also related to that of its component probabilities. Their interpretation comes from the basic components of association measures (see the definition in Section 1). A common approach is to estimate probabilities using relative frequency. Nominators of Φ components cover three of four basic values of the contingency table, i.e., a, b and c. The range of the d value is provided automatically because all probabilities have $n = a + b + c + d$ as denominators.

The component $p(xy)^{\alpha}$ represents the basic count of positively associated objects. It holds the elementary information we are usually looking for, and thus is the main component of most association measures. For a vast majority of association measures $\alpha \neq 0$, including all shown in Table 1. The higher the value of $\alpha > 0$, the larger prominence of objects with frequently associated features.

The component

$$p(x)^{\beta}p(y)^{\beta} = \left(\frac{a^2+ab+ac+bc}{n^2}\right)^{\beta}$$

holds symmetric information about objects sharing at least one feature. Objects sharing one and two features have similar influence on this component. If it is used as a normalization factor, we usually have $2\beta = -\alpha$. Association measures with asymmetric use of marginal probabilities have also been discussed in the literature, e.g., semi-conditional information (Washtell and Markert, 2009). Generic modeling of these measures is a separate topic and is not addressed here.

The last component is $\left(\frac{1}{2}\left(p(x)+p(y)\right)\right)^{\gamma}$ and it models the mean of the marginal probabilities. If it is used for normalization, we usually have $\gamma = -\alpha$. A more generic relation can be defined for this purpose, i.e., $\alpha = -\frac{1}{2}\beta - \gamma$. It is chosen as a compromise between modeling the sum of marginal probabilities and the probability of sum of events. Probability of sum of events gives fewer possibilities in association measure modeling. Two interesting cases are the Jaccard and second Kulczynski coefficients. The former requires the probability of a sum of events, but it also belongs to group T, which can be represented by the sum of marginal probabilities. The latter combines both the sum and the product of marginal probabilities (see Eqn. (64) in the next section).

3.2. Formulation of classic coefficients. Let us now define a set of association measure coefficients using the proposed one. The joint probability (group RR) is formulated as

$$\Phi(1, 0, 0) = p(xy). \tag{52}$$

In consequence (see Statement 1), equal values are also generated for the Russel–Rao coefficient:

$$\Phi(1, 0, 0) = \frac{a}{a+b+c+d} = RR. \tag{53}$$

The Consonni T3 coefficient ranking is generated by $\Phi(1, 0, 0)$ given that n is constant for all measured objects. The proof is identical to that for the Russel–Rao coefficient and is given in Statement 2. The Nieddu

coefficient with $\alpha = 0$, $\beta = 1$ and $\gamma = 1$ is also expressed by the same formula:

$$\Phi(1,0,0) = \frac{a}{a+b+c+d} = S_{NR}(0,1,1). \tag{54}$$

Pointwise mutual information (group F) is formulated as a monotone function of Φ. As a result, the Forbes and Gilbert–Wells coefficients may also be formulated. When collocation rankings are of interest, the logarithm in pairwise mutual information can be omitted:

$$\log \Phi(1,-1,0) = \log \frac{p(xy)}{p(x)p(y)} = PMI. \tag{55}$$

Mutual dependency (group DK) is formulated as

$$\log \Phi(2,-1,0) = \log \frac{p(xy)^2}{p(x)p(y)} = MD. \tag{56}$$

The Sorgenfrei coefficient does not require a logarithm and is defined as

$$\begin{aligned} \Phi(2,-1,0) &= \frac{p(xy)^2}{p(x)p(y)} \\ &= \frac{a^2}{(a+b)(a+c)} = Sorg. \end{aligned} \tag{57}$$

In consequence (see Statement 14), equal ranking is also generated for the Driver–Kroeber, Ochiai and Otsuka coefficients:

$$\begin{aligned} \Phi\left(1,-\frac{1}{2},0\right) &= \frac{p(xy)}{\sqrt{p(x)p(y)}} \\ &= \frac{a}{\sqrt{(a+b)(a+c)}} = DK. \end{aligned} \tag{58}$$

Log frequency biased mutual dependency (group L) is formulated as

$$\log \Phi(3,-1,0) = \log \frac{p(xy)^3}{p(x)p(y)} = fbMD. \tag{59}$$

Since n is constant for all observed objects, ranking equal to the frequency symmetric conditional probability (Buczyński, 2004) is generated as

$$\begin{aligned} \frac{1}{n}\Phi(3,-1,0) &= \frac{1}{n}\frac{p(xy)^3}{p(x)p(y)} \\ &= \frac{f(xy)^3}{f(x)f(y)} = FSCP. \end{aligned} \tag{60}$$

The Dice coefficient is formulated using the following equation:

$$\Phi(1,0,-1) = \frac{2p(xy)}{p(x)+p(y)} = \frac{2a}{2a+b+c} = D. \tag{61}$$

In consequence, all other coefficients of group T (Jaccard, normalized expectation, first Kulczynski, second Sokal–Sneath, Anderberg) are also generalized by $\Phi(1,0,-1)$. The Bray–Curtis coefficient also belongs to group T. It is an anti-monotone transformation of the Dice coefficient. Taking into account the property given by Eqn. (50), the association order of the Bray–Curtis coefficient is defined as

$$\begin{aligned} 1 - \frac{2}{\Phi(-1,0,1)} &= 1 - \frac{2}{\frac{p(x)+p(y)}{2p(xy)}} \\ &= 1 - \frac{2a}{2(a+b+c)} \\ &= \frac{b+c}{a+b+c} = BC. \end{aligned} \tag{62}$$

Mutual expectation is formulated as

$$\begin{aligned} \Phi(2,0,-1) &= \frac{2p(xy)^2}{p(x)+p(y)} \\ &= \frac{2f(xy)}{f(x)+f(y)}p(xy) = ME. \end{aligned} \tag{63}$$

The association order of the second Kulczynski coefficient is defined as

$$\begin{aligned} &\Phi(1,-1,1) \\ &= \frac{p(xy)(p(x)+p(y))}{2p(x)p(y)} \\ &= \frac{\frac{1}{n^2}a(a+b+a+c)}{\frac{2}{n^2}(a+b)(a+c)} = \frac{a(a+b+a+c)}{2(a+b)(a+c)} \\ &= \frac{a(a+c)}{2(a+b)(a+c)} + \frac{a(a+b)}{2(a+b)(a+c)} \\ &= \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right) = K_2. \end{aligned} \tag{64}$$

3.3. Formulation of generalized coefficients. The proposed coefficient Φ also generalizes coefficient rankings of other generalized coefficients. Relations between coefficients are visually shown in Fig. 6.

Statement 23. *The symmetric Tversky ($\alpha = \beta$) and Gover T_θ coefficients are monotone functions of the coefficient Φ. The Nieddu coefficient with $\alpha = \beta = 0$ is also generalized by the coefficient Φ. They all generate identical rankings of associations.*

Proof. Let $\theta \neq 1/2$. Then

$$\begin{aligned}
&\frac{1}{2\theta-1}\left[\frac{2\theta}{2\theta-(2\theta-1)\Phi(1,0,-1)}-1\right]\\
&=\frac{1}{2\theta-1}\left[\frac{2\theta}{2\theta-(2\theta-1)\frac{2a}{2a+b+c}}-1\right]\\
&=\frac{1}{2\theta-1}\left[\frac{2\theta}{\frac{2a+2\theta b+2\theta c}{2a+b+c}}-1\right]\\
&=\frac{1}{2\theta-1}\left[\frac{2\theta a+\theta b+\theta c}{a+\theta b+\theta c}-1\right]\\
&=\frac{1}{2\theta-1}\frac{(2\theta-1)a}{a+\theta b+\theta c}=\frac{a}{a+\theta b+\theta c}=T_\theta.
\end{aligned} \tag{65}$$

Let $\theta = 1/2$. Then

$$\Phi(1,0,-1)=\frac{a}{a+\frac{1}{2}b+\frac{1}{2}c}=T_{\frac{1}{2}}. \tag{66}$$

■

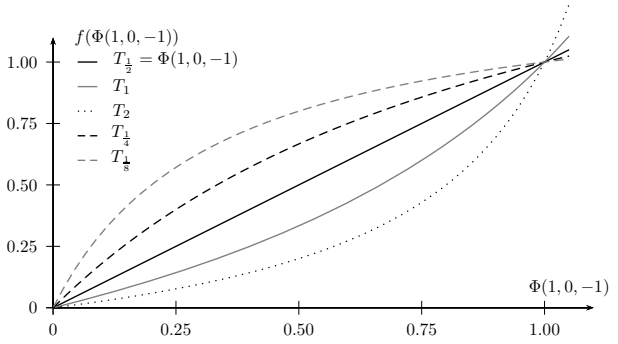

Fig. 6. Relations between Gover T_θ and the proposed coefficient $\Phi(1,0,-1)$.

To sum up, using the proposed coefficient Φ, we are able to generate rankings equal to at least 20 of widely known and used collocation association measures. Association rankings generated by the coefficients of groups RR, T, DK, F and L are covered. Another advantage of Φ is the ability to generate both normal and reversed orders of association, due to the property given by Eqn. (50). A summary of coefficients generated by the generalized coefficient Φ is given in Table 1.

4. Conclusion

The paper presented a detailed analysis of a set of binary association coefficients. The analysis focused on the association order equivalence between these coefficients. A generalized coefficient was proposed, able to generate ranking equivalent to at least 20 of association coefficients. Additionally, by negating parameters of the coefficient, reversed rankings of all association measures can be generated.

The work has its background in machine learning, pattern recognition and feature selection. Usage of presented association coefficients is common in many machine learning applications. Well designed machine learning methods usually work better on short, non-repeating sets of features. The main reasons and typical problems include the *curse of dimensionality*, the zero *generalized variance* or a *near-singular covariance matrix* caused by numerical representation errors. Introduction of too many features generates unnecessary noise in the data and makes the approach unclear. Additionally, some presented coefficients are anti-monotone transformations of others. Simultaneous usage of a coefficient and its anti-monotone transformation as features may be considered controversial from a machine learning point of view.

The main goal of this paper was to overview and simplify the usage of binary association measures. New association coefficients constantly appear. Some of them have interesting properties and are worth of interest. We encourage other researchers to seek new relations as these coefficients are proposed.

Acknowledgment

The author would like to thank Dr. Maciej Piasecki and Prof. Halina Kwaśnicka for their critical review of this paper. The present work has been financed within the *Common Language Resources and Technology Infrastructure (CLARIN-PL)* project.

References

Batagelj, V. and Bren, M. (1995). Comparing resemblance measures, *Journal of Classification* **12**(1): 73–90.

Buczyński, A. (2004). *Text Acquisition from the Internet for Linguistic Research*, Master's thesis, Warsaw University, Warsaw, (in Polish).

Chapelle, O. and Wu, M. (2010). Gradient descent optimization of smoothed information retrieval metrics, *Information Retrieval* **13**(3): 216–235.

Cheetham, A.H. and Hazel, J.E. (1969). Binary (presence-absence) similarity coefficients, *Journal of Paleontology* **43**(5): 1130–1136.

Choi, S.-S., Cha, S.-H. and Tappert, C.C. (2010). A survey of binary similarity and distance measures., *Journal of Systemics, Cybernetics & Informatics* **8**(1): 43–48.

Clarke, K.R., Somerfield, P.J. and Chapman, M.G. (2006). On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray–Curtis coefficient for denuded assemblages, *Journal of Experimental Marine Biology and Ecology* **330**(1): 55–80.

Consonni, V. and Todeschini, R. (2012). New similarity coefficients for binary data, *Match-Communications in Mathematical and Computer Chemistry* **68**(2): 581.

Dice, L.R. (1945). Measures of the amount of ecologic association between species, *Ecology* **26**(3): 297–302.

Duarte, J.M., Santos, J.B.d. and Melo, L.C. (1999). Comparison of similarity coefficients based on RAPD markers in the common bean, *Genetics and Molecular Biology* **22**(3): 427–432.

Friedman, J.H. (1997). On bias, variance, 0/1–loss, and the curse-of-dimensionality, *Data Mining and Knowledge Discovery* **1**(1): 55–77.

Gower, J.C. and Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients, *Journal of Classification* **3**(1): 5–48.

Hoang, H.H., Kim, S.N. and Kan, M.-Y. (2009). A re-examination of lexical association measures, *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications, Singapore*, pp. 31–39.

Hubalek, Z. (1982). Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation, *Biological Reviews* **57**(4): 669–689.

Jaccard, P. (1912). The distribution of the flora in the alpine zone 1, *New Phytologist* **11**(2): 37–50.

Johnson, R.A. and Wichern, D.W. (2007). *Applied Multivariate Statistical Analysis*, 6th Edn., Pearson International Edition, Prentice Hall, Upper Saddle River, NJ.

Kazienko, P. (2009). Mining indirect association rules for web recommendation, *International Journal of Applied Mathematics and Computer Science* **19**(1): 165–186, DOI: 10.2478/v10006-009-0015-5.

Kekäläinen, J. (2005). Binary and graded relevance in IR evaluations—comparison of the effects on ranking of IR systems, *Information Processing & Management* **41**(5): 1019–1033.

Liu, T.-Y. (2009). Learning to rank for information retrieval, *Foundations and Trends in Information Retrieval* **3**(3): 225–331.

Nieddu, L. and Rizzi, A. (2007). Proximity measures in symbolic data analysis, *Statistica* **63**(2): 195–211.

Pecina, P. (2005). An extensive empirical study of collocation extraction methods, *Proceedings of the Association for Computational Linguistics Student Research Workshop, Ann Arbor, MI, USA*, pp. 13–18.

Pecina, P. (2008). A machine learning approach to multiword expression extraction, *Proceedings of the Language Resources and Evaluation Workshop Towards a Shared Task for Multiword Expressions, Marrakech, Morocco*, pp. 54–61.

Pecina, P. (2010). Lexical association measures and collocation extraction, *Language Resources and Evaluation* **44**(1–2): 137–158.

Pecina, P. and Schlesinger, P. (2006). Combining association measures for collocation extraction, *Proceedings of the COLING/Association for Computational Linguistics on Main Conference, Sydney, Australia*, pp. 651–658.

Petrović, S., Šnajder, J. and Bašić, B.D. (2010). Extending lexical association measures for collocation extraction, *Computer Speech & Language* **24**(2): 383–394.

Rifqi, M., Lesot, M.-J. and Detyniecki, M. (2008). Fuzzy order-equivalence for similarity measures, *Annual Meeting of the North American Fuzzy Information Processing Society, NAFIPS 2008, New York, NY, USA*, pp. 1–6.

Segond, M. and Borgelt, C. (2011). Item set mining based on cover similarity, *in* J.Z. Huang, L. Cao and J. Srivastava (Eds.), *Advances in Knowledge Discovery and Data Mining*, Springer, Berlin/Heidelberg, pp. 493–505.

Tan, P.-N., Kumar, V. and Srivastava, J. (2004). Selecting the right objective measure for association analysis, *Information Systems* **29**(4): 293–313.

Tversky, A. (1977). Features of similarity, *Psychological Review* **84**(4): 327.

Washtell, J. and Markert, K. (2009). A comparison of windowless and window-based computational association measures as predictors of syntagmatic human associations, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore*, Vol. 2, pp. 628–637.

Wolda, H. (1981). Similarity indices, sample size and diversity, *Oecologia* **50**(3): 296–302.

Mariusz Paradowski graduated in computer science from the Wrocław University of Technology, Poland. In 2008 he received his Ph.D. in computer science with honors from the Institute of Informatics, and from 2009 to 2010 he was a research fellow at the Nanyang University of Technology, Singapore. Currently, he is an assistant professor at the Wrocław University of Technology. His research interests include theoretical and practical machine learning, artificial intelligence and computer vision.

Received: 22 April 2014
Revised: 21 August 2014
Re-revised: 30 September 2014