Silesian J. Pure Appl. Math.
vol. 6, is. 1 (2016), 69–83

Piotr WILCZEK

Computer Laboratory, Poznań, Poland

THE ULTRAMETRIC PROPERTIES OF BINARY DATASETS

Abstract. Many multivariate algorithms commonly applied for binary datasets depend on a proper metric (i.e., dissimilarity function) imposed on binary vectors. In the following work the relationships between different metrics defined on the randomly generated binary datasets and the *cophenetic correlation coefficient* (CCC) will be presented.

1. Introduction

Many data mining models are build on some similarity (or dissimilarity) measure which is used in order to collate how similar/dissimilar two data elements are. In the modern multivariate analysis such techniques as non-metric dimensional scaling, principal coordinate analysis and cluster analysis are instances of studies that strongly depend on the adequate choice of similarity/dissimilarity measure [15, 21, 24].

In the case of clustering it can be observed that the widespread applications of this method have shown a number of problems encountered during cluster analysis. In the data mining literature it is emphasized that the objective nature of cluster analysis is compromised by the subjective selections of clustering algorithm and measure of similarity/dissimilarity, since both the method and the measure exert significant impact on the analytical outcome [15, 24].

2010 Mathematics Subject Classification: 68P05, 91C20.
Keywords: binary data sets, dissimilarity function, clustering, cophenetic correlation coefficient.
Corresponding author: P. Wilczek (edwil@mail.icpnet.pl).
Received: 08.07.2016.

In the studies comparing similarity and dissimilarity coefficients applicable to binary strings it is widely recognized that the behavior of these indices is *data-specific* (i.e., dependent on the relative frequency of ones and zeros). Therefore, the choice of an adequate measure is largely subjective and often is based on tradition or on *a posteriori* criteria, for instance on the interpretability of the results. It can be asserted that the selection of an proper measure is essential in cluster analysis. For instance, Hastie, Tibshirami and Friedman stated in their comprehensive book "*The Elements of Statistical Learning*" that "[s]*pecifying an appropriate dissimilarity measure is far more important in obtaining success with clustering than the choice of clustering algorithm. This aspect of the problem is emphasized less in the clustering literature than the algorithms themselves, since it depends on domain knowledge specifies and is less amenable to general research*" [12].

Since the results of cluster analysis may depend on the choice of similarity/dissimilarity measure, we need to understand the behavior of different types of coefficients. Some indices have mathematical properties that make them inadequate for certain analyses. Also some kinds of clustering techniques will only yield reproducible and meaningful results when certain indices are employed. The implications of the selection of clustering method are well established [15, 24]. On the other hand, the implications of choosing different similarity/dissimilarity coefficients in the case of cluster analysis of binary data are vaguely recognized. There are numerous studies available which evaluate the usefulness of different similarity/dissimilarity measures for clustering of continuous data, but there are only few analyses comparing the influence of different indices on the results of clustering procedures applied to categorical (e.g., binary) datasets [22, 23]. Furthermore, all studies concerning binary datasets are confined only to low dimensional sets and are restricted to the domain-specific data. Hence, it seems valuable to undertake the comparative analysis of diverse binary similarity/dissimilarity measures in the case of low as well as high dimensional datasets.

Clustering of binary data has attracted increasing attention since a variety of natural phenomena as well as artifacts can be adequately described in the terms of bivariate sequences. Also many multivariate processes can be modelled by the dichotomized data. Such instances as standard Boolean model in information retrieval [11], Galois lattices in formal concept analysis [8], two-mode social networks in network sciences [28], transactional databases [13, 18, 20], bivalued logical matrices in quantum theory [29, 30], presence/absence matrices in ecological studies and many others straightforwardly indicate that binary datasets are ubiquitous in natural, technical and social sciences [25].

Summing up, it can be claimed that the objective of the present studies was to compare and benchmark the alternative similarity and dissimilarity measures for clustering of binary datasets, to determine the relatedness between these indices and to scrutinize their usefulness while they are applied to low and high dimensional datasets as well as to sparse and dense binary datasets.

The random sampling procedure has enabled us to arrive at some conclusions about the behavior of seven investigated similarity/dissimilarity coefficients in the above contexts.

In this paper we will identify a binary dataset with a Boolean matrix (of size $k \times n$) whose rows are indexed by the set $O = \{x_1, x_2, \ldots, x_k\}$ of *objects* (or synonymously, *data points*, *instances*, *cases*, *persons*, *entities*, *patterns*, *tuples*, *transactions*) and whose columns are indexed by the set $A = \{a_1, a_2, \ldots, a_n\}$ of *attributes* (or synonymously, *features*, *variables*, *items*, *events*, *dimensions*, *components*). A Boolean row vector of dimension n is an $n-$tuples $x = \{b_1, b_2, \ldots, b_n\}$ where $b_i \in \{0, 1\}$. The set of all Boolean vectors of dimension n is denoted by $\mathcal{B}_n$. The characteristic vector of a subset ($A' \subseteq A$) of the attributes is the Boolean row vector $x(A') \in \mathcal{B}_n$ whose $i-$th component is equal to 1 if and only if $a_i \in A'$ and is equal to 0 if and only if $a_i \notin A'$. We will identify the set $\mathcal{B}_n$ with $n-$dimensional dataset on which different functions can be defined.

2. The dissimilarities on $\mathcal{B}_n$ and the cophenetic correlation coefficient

One of the most popular approach in comparing two equal-length binary vectors is to introduce some kind of similarity coefficient in order to quantitatively assess the resemblance between them. Customary it is assume that any similarity coefficient s defined on the dataset $\mathcal{B}_n$ should satisfy the conditions:

a) $s(x_i, x_j) = s(x_j, x_i)$,

b) $s(x_i, x_i) = 1$ and

c) $s(x_i, x_j) \leqslant 1$ for all $x_i, x_j \in \mathcal{B}_n$ [9,10].

Similarity coefficients which were primarily proposed for binary data such as presence/absence matrices or answers to yes-no questions attain their maxima when two patterns possess identical values across their variables. It should be pointed out that similarity indices are never metric since it is always possible to indicate

two datapoints, A and B, that are more similar than the sum of their resemblances with another datapoint C.

In analogy to the conditions imposed on the similarity coefficients there are axioms usually applied for the notion of dissimilarity. A function d is said to be a dissimilarity coefficient if it satisfies:

a) $d(x_i, x_j) = d(x_j, x_i)$,

b) $d(x_i, x_i) = 0$ and

c) $d(x_i, x_j) \geqslant 0$ for all $x_i, x_j \in \mathcal{B}_n$.

Dissimilarities measures which can be conceptually (and often also mathematically) understood as the complements of similarity indices reach their maxima when two datapoints share no common variable values. Further, we will identify the dataset $\mathcal{B}_n$ with the similarity s or dissimilarity function d defined on it with the $n-$dimensional similarity data space $(\mathcal{B}_n, s)$ or with $n-$dimensional dissimilarity data space $(\mathcal{B}_n, d)$, respectively. It is generally assumed that the structures $(\mathcal{B}_n, s)$ and $(\mathcal{B}_n, d)$ are completely characterized by their similarity or dissimilarity matrices [9,10]. Further, additional conditions can be imposed on the dissimilarity data space $(\mathcal{B}_n, d)$. Namely, if the function d also satisfies the requirement of the definiteness (i.e., $d(x_i, x_j) = 0$ if and only if $x_i = x_j$) and the triangle inequality, then d is said to be a metric. The dissimilarity data space $(\mathcal{B}_n, d)$ is said to be Euclidean if its datapoints can be embedded in a Euclidean space such that the Euclidean distance between $x_i, x_j \in \mathcal{B}_n$ is equal to $d(x_i, x_j)$ [10].

A direct manner to convert a similarity function s into a dissimilarity d is to take the complement $d = 1 - s$. But as pointed out by Gower and Legendre such obtained dissimilarity not always is a metric or not always has a Euclidean representation. They proved that if $(\mathcal{B}_n, s)$ is given by a positive semidefinite similarity matrix with elements $0 \leqslant s(x_i, x_j) \leqslant 1$ and $s(x_i, x_i) = 1$, then the dissimilarity data space $(\mathcal{B}_n, d)$ given by the dissimilarity matrix with elements defined as $d(x_i, x_j) = \sqrt{(1 - s(x_i, x_j))}$ is Euclidean (cf. Theorem 6 in [10]).

There exist many different similarity and dissimilarity indices [2, 3, 9, 10]. To formally characterize these coefficients let us introduce the following four terms: for any two datapoints $x_i, x_j \in \mathcal{B}_n$ the term $x_i^T x_j$ denotes the number of positions where both vectors x_i and x_j have value equal to 1, the term $\overline{x_i}^T \overline{x_j}$ denotes the number of positions where both patterns x_i and x_j have value equal to 0 (here, $\overline{x}$ is the complement of binary vector x defined as $\overline{x} = \mathbf{I} - x$, where $\mathbf{I}$ is the unit binary vector of the same dimension as x and x^T denotes the transposition of the

vector x). The terms $x_i^T\overline{x_j}$ and $\overline{x_i}^T x_j$ denotes the number of positions where the pattern x_i has value equal to 1 and the pattern x_j has value equal to 0 and the number of positions where x_i has value equal to 0 and x_j has the value equal to 1, respectively.

One of the most widely used measures in comparing two equal-length binary vectors is the Hamming dissimilarity d_H defined as:

$$d_H(x_i, x_j) = x_i^T\overline{x_j} + \overline{x_i}^T x_j$$

for all datapoints $x_i, x_j \in \mathcal{B}_n$. This index counts the number of bits that are different in two patterns x_i and x_j. It was shown that the Hamming dissimilarity is a metric [6]. d_H can be simply considered as geometrical L_1 distance applied to $n-$dimensional binary dataset $\mathcal{B}_n$. This metric has not Euclidean representation. The so-called Hamming similarity s_H is the number of identical positions in two strings x_i and x_j, i.e.,

$$s_H(x_i, x_j) = x_i^T x_j + \overline{x_i}^T\overline{x_j}.$$

The coefficient s_H does not satisfy the conditions from the theorem of Gower and Legendre and can not be directly transformed into an Euclidean dissimilarity index. But this similarity measure was normalized by Sokal and Michener to the form of the so-called simple matching coefficient s_{SM} which is defined as:

$$s_{SM}(x_i, x_j) = \frac{x_i^T x_j + \overline{x_i}^T\overline{x_j}}{x_i^T x_j + x_i^T\overline{x_j} + \overline{x_i}^T x_j + \overline{x_i}^T\overline{x_j}}$$

for all $x_i, x_j \in \mathcal{B}_n$. Another normalization of s_H was introduced by Rogers and Tanimoto:

$$s_{RT}(x_i, x_j) = \frac{x_i^T x_j + \overline{x_i}^T\overline{x_j}}{x_i^T x_j + 2(x_i^T\overline{x_j} + \overline{x_i}^T x_j) + \overline{x_i}^T\overline{x_j}}$$

for all datapoints $x_i, x_j \in \mathcal{B}_n$. Therefore, The Sokal-Michener simple matching coefficient s_{SM} as well as the Rogers-Tanimoto similarity measure s_{RT} are considered as the Hamming-based similarities. These two similarity functions can be transformed according to the rule $d_z(x_i, x_j) = \sqrt{(1 - s_z(x_i, x_j))}$ for $z \in \{SM, RT\}$ in order to yield corresponding Euclidean dissimilarities [10].

Another similarity measure is the *Phi* coefficient which is defined as:

$$s_{Phi} = \frac{x_i^T x_j \times \overline{x_i}^T\overline{x_j} - x_i^T\overline{x_j} \times \overline{x_i}^T x_j}{\sqrt{(x_i^T x_j + x_i^T\overline{x_j})(x_i^T x_j + \overline{x_i}^T x_j)(\overline{x_i}^T\overline{x_j} + x_i^T\overline{x_j})(\overline{x_i}^T\overline{x_j} + \overline{x_i}^T x_j)}}$$

for all $x_i, x_j \in \mathcal{B}_n$. This index is equivalent to the Pearson's product moment correlation applied to binary data. From this similarity function it is possible to obtain the following Euclidean dissimilarity measure [10]:

$$d_{Phi}(x_i, x_j) = \sqrt{(1 - s_{Phi}(x_i, x_j))}.$$

Recall that the term $x_i^T x_j$ is the inner product (IP) of two vectors x_i and x_j and it gives rise to the inner product similarity measure s_{IP} which has the form:

$$s_{IP}(x_i, x_j) = x_i^T x_j$$

for all $x_i, x_j \in \mathcal{B}_n$. This similarity function also does not satisfy the requirements of the theorem of Gower and Legendre and it is not possible to gain from s_{IP} an Euclidean dissimilarity index. But from this similarity index it is feasible to derive the following normalized similarity coefficients:

a) Jaccard similarity coefficient s_J:

$$s_J(x_i, x_j) = \frac{x_i^T x_j}{x_i^T x_j + x_i^T \overline{x_j} + \overline{x_i}^T x_j},$$

b) Dice-Sorensen similarity coefficient s_{DS}:

$$s_{DS}(x_i, x_j) = \frac{2x_i^T x_j}{2x_i^T x_j + x_i^T \overline{x_j} + \overline{x_i}^T x_j},$$

c) Ochia similarity coefficient s_O:

$$s_O(x_i, x_j) = \frac{x_i^T x_j}{\sqrt{(x_i^T x_j + x_i^T \overline{x_j})(x_i^T x_j + \overline{x_i}^T x_j)}}$$

for all binary patterns $x_i, x_j \in \mathcal{B}_n$.

These normalized similarity indices can be converted into the corresponding Euclidean dissimilarities according to the rule: $d_z(x_i, x_j) = \sqrt{(1 - s_z(x_i, x_j))}$ for $z \in \{J, DS, O\}$ [10].

The Hamming, Simple Matching Coefficient, Rogers-Tanimoto, Jaccard, Dice-Sorensen and Ochia dissimilarity indices are regarded as the so-called measures of co-occurrence whereas the Phi coefficient is the measure of association.

In order to illustrate the above introduced concepts let us consider two binary strings: $x_1 = (1,1,0,0,1,1,0,1,0)$ (whose complement is given by $\overline{x}_1 = (0,0,1,1,0,0,1,0,1)$) and $x_2 = (1,1,0,1,0,0,0,1,1)$ (whose complement is given by $\overline{x}_2 = (0,0,1,0,1,1,1,0,0)$). Then the four terms $x_1^T x_2$, $\overline{x}_1^T \overline{x}_2$, $x_1^T \overline{x}_2, \overline{x}_1^T x_2$ have the values equal to: 3, 2, 2 and 2, respectively. Consequently, the values of all similarities considered here are given by: $s_H(x_1,x_2) = 5$, $s_{SM}(x_1,x_2) = \frac{5}{9}$, $s_{RT}(x_1,x_2) = \frac{5}{13}$, $s_{Phi}(x_1,x_2) = \frac{1}{10}$, $s_{IP}(x_1,x_2) = 3$, $s_J(x_1,x_2) = \frac{3}{7}$, $s_{DS}(x_1,x_2) = \frac{3}{5}$, $s_O(x_1,x_2) = \frac{3}{5}$.

The Hamming dissimilarity $d_H(x_1,x_2)$ has the value equal to 4 and all dissimilarities obtainable from the corresponding similarity measures according to the transformation $d(x_i,x_j) = \sqrt{1 - s(x_i,x_j)}$ have the values: $d_{SM}(x_1,x_2) = 0.667$, $d_{RT}(x_1,x_2) = 0.784$, $d_{Phi}(x_1,x_2) = 0.949$, $d_J(x_1,x_2) = 0.756$, $d_{SD}(x_1,x_2) = 0.632$ and $d_O(x_1,x_2) = 0.632$, respectively.

If a dissimilarity d defined on $\mathcal{B}_n$ also satisfies the so-called strong triangle inequality (i.e., the requirement $d(x_i,x_j) \leqslant \max\{d(x_i,x_k), d(x_j,x_k)\}$ for any distinct $x_i, x_j, x_k \in \mathcal{B}_n$), then the structure $(\mathcal{B}_n, d)$ is an ultrametric data space. We will denote any ultrametric by d_u. In this case the pairwise distances between datapoints are given by an ultrametric (also known as cophenetic) distance matrix [15, 24].

Based on the assumption that all hierarchical clustering algorithms can be identified with a mappings from $n-$dimensional dissimilarity data space (X,d) into $n-$dimensional ultrametric data space (X,d_u) it can be observed that there exists a natural bijection between the set of all ultrametrics defined on the arbitrary dataset X and the set of all dendrograms defined on it [15,24]. Also it was proved that the single-linkage hierarchical clustering algorithm which take as an input the dissimilarity data space (X,d) returns the ultrametric data space (X,d_u) (i.e., dendrogram) such that the function d_u is the so-called maximal subdominant ultrametric for the dissimilarity d [24]. To define this concept formally, note first that the whole set of dissimilarities $\mathcal{D}$ defined on X can naturally be ordered as follows:

$$d' \leqslant d \longleftrightarrow d'(x_i,x_j) \leqslant d(x_i,x_j) \text{ for all } (x_i,x_j) \in X \times X \text{ and all } d, d' \in \mathcal{D}.$$

Then it can be shown that for any set $\mathcal{U}^{\downarrow} = \{d'_u \in \mathcal{U} \,|d'_u \leqslant d\}$ of all ultrametrics on X which are smaller than d there exists the maximal element $d_u^{\downarrow}$ of $\mathcal{U}^{\downarrow}$ which is unique. Such characterized ultrametric $d_u^{\downarrow}$ on X is termed the maximal subdominant ultrametric associated with the dissimilarity d.

In [26] Sokal and Rohlf introduced the so-called cophenetic correlation coefficient (CCC) in order to quantitatively evaluate how faithfully the ultrametric data space (X, d_u) reflects the pairwise distances between the datapoints from the dissimilarity data space (X, d). This coefficient is defined as the Pearson's product moment correlation coefficient between the dissimilarity matrix associated with (X, d) and the ultrametric distance matrix associated with (X, d_u):

$$c = \frac{\sum\limits_{i<j} \left(d(x_i, x_j) - \overline{d}\right) \left(d_u(x_i, x_j) - \overline{d_u}\right)}{\sqrt{\left[\sum\limits_{i<j} \left(d\left(x_i, x_j\right) - \overline{d}\right)^2\right] \left[\sum\limits_{i<j} \left(d_u\left(x_i, x_j\right) - \overline{d_u}\right)^2\right]}},$$

where $\overline{d}$ is the average dissimilarity in (X, d) and $\overline{d_u}$ is the average ultrametric distance in (X, d_u) [17, 21, 25, 26]. The high value of CCC indicates that the resulting dendrogram exhibits the low distortion of the original data. It is usually assumed that the clustering algorithm adequately represents the dissimilarity data space (X, d) if it produces the dendrograms (i.e., the ultrametric data spaces (X, d_u)) with the value of CCC equal or higher than 0.6 [1].

In our studies we will employ these general results to the dissimilarity data spaces of the form $(\mathcal{B}_n, d)$ and we will show how the dimension of $(\mathcal{B}_n, d)$, kind of dissimilarity d and the density of dataset $\mathcal{B}_n$ influence on the values of the cophenetic correlation coefficient for the ultrametric data spaces $(\mathcal{B}_n, d_u)$ obtained as the output of the single-linkage hierarchical clustering algorithm.

3. Experimental results and discussion

Assessing the adequacy of alternative similarity/dissimilarity measures for hierarchical clustering necessitates the crucial step of choosing reference datasets. A perfect reference dataset should imitate the variability encountered in experimental data and it should possess some *a priori* known structure in order to determine the appropriateness of the outcomes recorded from the alternative studies. Usually the methodologies based on parametric simulations, exampler datasets and permutation reshuffling have been used to obtain such reference datasets in assessing the adequacy of similarity/dissimilarity measures in clustering analysis.

To overcome the aforementioned data-specificity of the alternative coefficients it was decided to use the random sampling method in order to obtain the variety of reference datasets with some *a priori* determined structures. We have used the

randomly generated datasets whose dimensions ranged from 1000 to 100000 and whose densities ranged from 0.01 to 0.99.

Specifically, our methodology was based on the synthetic random bipartite graphs since it is well known that binary object-by-attribute datasets can be modelled by bipartite graphs [16, 19, 27]. Namely, every Boolean $k \times n$ matrix representing binary dataset with $|X| = k$ objects and $|A| = n$ attribute corresponds to the incidence matrix of bipartite graph $G = (X \cup A, E)$ such that there exists an edge $e \in E$ between vertices $x \in X$ and $a \in A$ if and only if the object x has the attribute a. Furthermore, it is possible to identify the density of a dataset with the edge-density of bipartite graph G given by the equation:

$$density(G) = \frac{m}{kn},$$

where $|E| = m$. Based on the above a one-to-one correspondence between binary datasets and bipartite graphs we have generated such graphs (of changing edge-density) using the random graph generation model of Erdos-Renyi of type $G(N, m)$, where $N = |X| + |A|$ and where the m edges are chosen uniformly randomly from the set of all possible edges between X and A [7]. We have chosen the number of objects $|X| = 50$ and the number of dimensions $|A| = 1000$, 10000, 20000, 30000, 40000, 50000, 60000, 70000, 80000, 90000, 100000. The parameter m was monitored in order to obtain the bipartite graphs corresponding to binary datasets with densities: 0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 0.99 in all dimensions. All datasets $\mathcal{B}_n$ have been endowed with the metric dissimilarity function d_z where $z \in \{H, SM, RT, J, DS, O, Phi\}$. Consequently, we have obtained 1617 dissimilarity data spaces of the form $(\mathcal{B}_n, d_z)$. All such created metric structures were subjected to the single-linkage hierarchical clustering algorithm in order to gain for each dissimilarity function d_z its corresponding maximal subdominant ultrametric. Furthermore, we have calculated the value of the cophenetic correlation coefficient for such formed ultrametric data spaces to assess quantitatively the distortion between the input dissimilarity index d_z and the output dendrograms (cf. Table 1).

All simulations and calculations were done in the R programming language [4, 5, 14].

In our studies it was shown that the Euclidean dissimilarities d_{SM} and d_{RT} yielded almost identical values of CCC as the non-Euclidean dissimilarity coefficient d_H for a given density of the dataset $\mathcal{B}_n$. These observations were independent of the dimension of $\mathcal{B}_n$. This behavior of the Hamming, the Sokal-Michener

Table 1
The values of the cophenetic correlation coefficient for seven dissimilarity data spaces for dimension $n = 100000$; the values above 0.6 are in bold

Density	$(\mathcal{B}_n, d_H)$	$(\mathcal{B}_n, d_{SM})$	$(\mathcal{B}_n, d_{RT})$	$(\mathcal{B}_n, d_J)$	$(\mathcal{B}_n, d_{DS})$	$(\mathcal{B}_n, d_O)$	$(\mathcal{B}_n, d_{Phi})$
0.01	**0.791**	**0.791**	**0.791**	0.258	0.258	0.258	0.262
0.05	**0.757**	**0.757**	**0.757**	0.143	0.143	0.142	0.170
0.1	**0.664**	**0.663**	**0.664**	0.260	0.259	0.259	0.162
0.15	**0.627**	**0.627**	**0.628**	0.213	0.213	0.212	0.153
0.2	**0.625**	**0.625**	**0.625**	0.225	0.225	0.225	0.207
0.25	0.430	0.431	0.431	0.227	0.227	0.227	0.129
0.3	0.265	0.265	0.266	0.253	0.252	0.252	0.225
0.35	0.271	0.271	0.272	0.402	0.402	0.401	0.171
0.4	0.288	0.288	0.289	0.335	0.334	0.333	0.176
0.45	0.193	0.193	0.193	0.402	0.402	0.401	0.191
0.5	0.204	0.205	0.205	0.241	0.241	0.241	0.205
0.55	0.239	0.239	0.240	0.423	0.423	0.422	0.235
0.6	0.240	0.241	0.241	0.484	0.484	0.483	0.216
0.65	0.280	0.280	0.281	0.459	0.459	0.458	0.178
0.7	0.335	0.335	0.335	0.483	0.483	0.483	0.173
0.75	0.446	0.446	0.446	**0.608**	**0.608**	**0.608**	0.176
0.8	**0.621**	**0.621**	**0.621**	**0.704**	**0.704**	**0.704**	0.120
0.85	0.575	0.575	0.575	**0.628**	**0.628**	**0.627**	0.224
0.9	**0.750**	**0.750**	**0.750**	**0.767**	**0.767**	**0.767**	0.247
0.95	**0.763**	**0.763**	**0.763**	**0.769**	**0.769**	**0.769**	0.202
0.99	**0.873**	**0.873**	**0.873**	**0.873**	**0.873**	**0.873**	0.191

and the Rogers-Tanimoto dissimilarities can be explained by the fact that these indices are symmetrical with respect to the values given by the four terms: $x_i^T x_j$, $\overline{x_i}^T \overline{x_j}$, $x_i^T \overline{x_j}$, $\overline{x_i}^T x_j$. The metric d_H excludes simultaneously the joint presence of 0 and 1 in both vectors $x_i, x_j \in \mathcal{B}_n$ whereas the coefficients d_{SM} and d_{RT} include all four terms. Recall that the Simple Matching Coefficient of Sokal and Michener gives equal weight to both forms of agreement between two binary patterns, i.e. double zeros and double ones. Therefore, when this index is used it is *implicitly* assumed that there is no difference between positive matches and negative matches. In turn, the Rogers-Tanimoto coefficient give double weight to mismatches (i.e., $x_i^T \overline{x_j}$ and $\overline{x_i}^T x_j$).

The Hamming metric d_H, the Sokal-Michener metric d_{SM} and the Rogers-Tanimoto metric d_{RT} attained the satisfactory values of CCC (i.e., equal or higher than 0.6) only for the relatively sparse (i.e., for $density(\mathcal{B}_n) \leqslant 0.2$) and relatively dense (i.e., for $density(\mathcal{B}_n) \geqslant 0.8$) binary datasets. These regularities were noted regardless of the dimension of the dissimilarity data space.

The behavior of these indices can be presumably elucidated by the fact that the Hamming measure whose formula includes only mismatches (i.e., $x_i^T \overline{x_j}$ and

$\overline{x_i}^T x_j$), the Simple Matching Coefficient and the Rogers-Tanimoto measure whose formulae include joint absences in the numerators initiate clusters from the vectors with the high frequency of ones as well as with the high frequency of zeros. The simultaneous exclusion of double presences and double absences (in the case of d_H) or the inclusion of joint double zeros (in the case of d_{SM} and d_{RT}) provides equal importance to positive as well as negative matches. Therefore, the strings with many 1 are as important as the strings with a few 1 in cluster formation.

Another tested dissimilarity measure was the Phi measure. The Phi metric also is symmetrical but – in contrast with the Hamming-based coefficients d_{SM} and d_{RT} – this distance measure can be seen as the multiplicative form of the terms $x_i^T x_j$ and $\overline{x_i}^T \overline{x_j}$ whereas the Sokal-Michener and the Rogers-Tanimoto dissimilarities are additive forms of these terms. Nevertheless, the influence of positive and negative matches are treated equally important in the Phi index as well as Hamming-based measures. The numerator of the Phi measure is the determinant of the 2×2 contingency table for two dichotomous variables. In fact, the Phi coefficient is the square root of the χ^2 (chi-square) statistics for 2×2 frequency tables. It was observed that in the Phi-based dendrogram neither vectors with the high frequency of ones nor vectors with the low frequency of ones are favored. Regardless of the density and dimension of $\mathcal{B}_n$ the values of CCC for the Phi metric never achieved the value higher than 0.35.

Also three Euclidean dissimilarities d_J, d_{DS}, d_O always yielded almost identical values of CCC for a given density of the dataset $\mathcal{B}_n$. It was observed for all dimensions of the data space. For the Jaccard metric d_J, the Dice-Sorensen metric d_{DS} and the Ochia metric d_O the sufficiently good values of CCC are reached only in the case of the datasets $\mathcal{B}_n$ of the density equal or higher than 0.75 (cf. Table 1). Recall that these three coefficient are symmetrical. The Jaccard measure excludes double zeros and gives equal weight to positive matches and to mismatches. On the other hand, the Sorensen-Dice measure is similar to d_J but it gives double weight to non-zero agreement between two strings. Thus, d_{DS} is monotonic with respect to d_J and if the resemblance for a pair of datapoints obtained with the Jaccard measure is higher than that of another pair of data elements then the same will be true when using the Sorensen-Dice coefficient. In the case of d_{DS} it can be ascertained that the co-occurrence of variables among two patterns is more informative or more important than their absences. Negative matches may be brought about by various factors. They do not necessarily reflect differences in two experimental data elements. On the contrary, double presences can be undoubtedly regarded as a strong indication of similarity. The Ochia dissimilarity is the quotient of the positive matches between two vectors and the square rooted

product of the sums of positive matches and each form of mismatches (i.e., $x_i^T\overline{x_j}$ and $\overline{x_i}^T x_j$). Whereby, this coefficient is based on the idea of the geometric mean and therefore outcomes with different ranges will be normalized before a resulting value is obtained. This index is especially suitable in the case when the ranges and variance of positive matches are very different from one another. Coefficients including only double positive matches in the numerator initiate clusters from the strings with the high frequency of ones and therefore the indices d_J, d_{DS}, d_O attained the highest values of CCC when applied to dense binary datasets.

Summing up the above results it can be asserted that the quantitative measure of the distortion of the transformation $d(x_i, x_j) \rightarrow d_u(x_i, x_j)$ (where d_u is the maximal subdominant ultrametric obtained as the result of the single-linkage hierarchical clustering algorithm) in the form of the cophenetic correlation coefficient depends strongly on the kind of the input dissimilarity function d as well as on the density of the underlying datasets $\mathcal{B}_n$.

4. Conclusions

The synthetic binary datasets of the different dimensions and of the different densities were used to analyze seven dissimilarity functions. Comparisons were made of Hamming, Simple Matching, Rogers-Tanimoto, Phi, Jaccard, Sorensen-Dice and Ochia measures using results from the distance-based clustering algorithm applied to binary datasets. All tested dissimilarities have Euclidean properties with the exception of the Hamming distance. The cophenetic correlation coefficient was calculated for all distance measures. Dendrograms obtained by the single-linkage clustering method have shown the high level of redundancy among the co-occurrence coefficients. The Hamming, Simple Matching and Rogers-Tanimoto measures have produced dendrograms with the nearly identical ultrametric matrices. This pattern is also easily seen from the Jaccard, Sorensen-Dice and Ochia dendrograms. This behavior of the co-occurence indices can be explained by the fact that the functions d_H, d_{SM} and d_{RT} are symmetrical with respect to the terms $x_i^T x_j$, $\overline{x_i}^T\overline{x_j}$, $x_i^T\overline{x_j}$, $\overline{x_i}^T x_j$. On the other hand, the Jaccard, Sorensen-Dice and Ochia indices are rendered as asymmetrical since they ignore the term $\overline{x}_i^T\overline{x}_j$ and – consequently – do not involve negative matches. It can be claimed that the inclusion or exclusion of double zeros has considerable influence on the structure of the resulting dendrograms since in the case of binary data such as presence/absence matrices zero values shared between two datapoints are

not necessarily indications of the resemblance between these data elements. Zeros values may arise not only in the cases of true absences of same variables but may simply indicate that some variables were not reported or measured. In symmetrical coefficients, the value zero for two data elements is considered in exactly the same manner as any other pair of values. These indices can be used when the state zero is an acceptable ground for comparing two data points and this state is regarded as the same sort of information as any other value. All six co-occurence indices perform differently for datasets with different densities. Also it was demonstrated that the *Phi*-based ultrametric matrices attain unsatisfactory level of CCC. Thus, the *Phi* distance should be used only in the special circumstances, in which the researcher has some strong justification for this dissimilarity.

Accordingly, these findings suggest that due to the indices with similar properties exhibit very similar outcomes, the choice between them should be based on the fact of considering or not the negative co-occurrences in their formulae.

Acknowledgment

The author would like to express his thanks to the anonymous reviewer for his valuable comments and suggestions.

References

1. Cassu C., Ferrer J.G., Bonet J.: *Classification of several business sectors according to uncertain characteristics*. In: Handbook of management under uncertainty. Gil-Aluja J. (ed.), Kluwer Academic Publ., Dordrecht 2001.
2. Cha S-H.: *Comprehensive survey on distance/similarity measures between probability density functions*. Int. J. Math. Models Methods Appl. Sci. **1** (2007), 300–307.
3. Choi S-S., Cha S-H, Tappert C.C.: *A survey of binary similarity and distance measures*. Journal of Systemics, Cybernetics and Informatics **8** (2010), 43–48.
4. Csardi G., Nepusz T.: *The igraph software package for complex network research*. InterJournal Complex Syst. **1695** (2006), http://igraph.org
5. Dray S., Dufour A.B.: *The ade4 package: implementing the duality diagram for ecologists*. J. Stat. Soft. **22** (2007), 1–20.
6. Duda R.O., Hart P.E., Stork D.G.: *Pattern Classification*. Wiley Interscience, New York 2000.

7. Erdos P., Renyi A.: *On random graphs*. Publ. Mat. **6** (1959), 290-297.
8. Ganter B., Wille R.: *Formal Concept Analysis: Mathematical Foundations*. Springer-Verlag, Berlin 1998.
9. Gower J.C.: *A general coefficient of similarity and some of its properties*. Biometrics **27** (1971), 857-871.
10. Gower J.C., Legendre P.: *Metric and Euclidean properties of dissimilarity coefficients*. J. Classification **3** (1986), 5-48.
11. Grossman D.A., Frieder O.: *Information Retrieval: Algorithms and Heuristics*. Springer, Dordrecht 2004.
12. Hastie T., Tibshirani R., Friedman J.: *Elements of Statistical Learning*. Springer-Verlag, New York 2009.
13. Ibrahim H.M., Marghny M.H., Abdelaziz N.M.A.: *Fast vertical mining using Boolean algebra*. Int. J. Adv. Comput. Sci. Appl. **6** (2015), 89–96.
14. Ihaka R., Gentleman R.: *A language for data analysis and graphics*. J. Comput. Graph. Statist. **5** (1996), 299–314.
15. Jain A.K., Dubes R.C.: *Algorithms for Clustering Data*. Prentice Hall, New Jersey 1988.
16. Jayalakshmi G., Nageswara Rao K.: *Mining association rules for large transactions using new support and confidence measures*. J. Theor. Appl. Inform. Technol. **7** (2009), 94–100.
17. Lapointe F-J., Legendre P.: *Statistical significance of the matrix correlation coefficient for comparing independent phylogenetic trees*. Systematic Biology **41** (1992), 378–384.
18. Li G., Zaki M.J.: *Sampling frequent and minimal Boolean patterns: theory and application in classification*. Data Min. Knowl. Discov. **30** (2016), 181–225.
19. Li J., Li H., Soh D., Wong L.: *A correspondence between maximal complete bipartite subgraphs and closed patterns*. Lecture Notes in Comput. Sci. **3721** (2005), 146–156.
20. Pathi S., Kothalanka A., Addala V.: *Binary matrix approach for mining frequent sequential pattern in large databases*. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **4** (2014), 311–317.
21. Podani J., Dickinson T.A.: *Comparison of dendrograms: a multivariate approach*. Can. J. Botany **62** (1984), 2765–2778.
22. Sesli M., Yegenoglu E. D.: *Comparison of similarity coefficients used for cluster analysis based on RAPD markers in wild olives*. Genet. Mol. Res. **9** (2010), 2248–2253.

23. Shirkhorshidi A.S., Aghabozorgi S., Wah T.Y.: *A comparison study on similarity and dissimilarity measures in clustering continuous data*. PloS ONE **10**(12): e0144059 (2015), 1–20.
24. Simovici D.A., Djeraba C.: *Mathematical Tools for Data Mining*. Springer-Verlag, London 2008.
25. Sneath P.H.A., Sokal R.R.: *Numerical Taxonomy. The Principles and Practice of Numerical Classification*. W.H. Freeman and Co., San Francisco 1973.
26. Sokal R.R., Rohlf F.J.: *The comparison of dendrograms by objective methods*. Taxon **9** (1962), 33–40.
27. Spyropoulou E., De Bie T., Boley M.: *Interesting pattern mining in multi – relational data*. Data Min. Knowl. Discov. **28** (2014), 808–849.
28. Wasserman S., Faust K.: *Social Network Analysis: Methods and Applications*. Cambridge Univ. Press, Cambridge 1994.
29. Wilczek P.: *Model-theoretic investigations into consequence operation (Cn) in quantum logics: an algebraic approach*. Internat. J. Theoret. Phys. **45** (2006), 679-689.
30. Wilczek P.: *Constructible models of orthomodular quantum logics*. Electron. J. Theor. Phys. **5** (2008), 9–32.