Metodološki zvezki, Vol. 17, No. 1, 2020, 33–54

Associations among similarity and distance measures for binary data in cluster analysis

Jana Cibulková[1] Zdenek Šulc[2] Hana Řezanková[3]
Sergej Sirota[4]

Abstract

The paper focuses on similarity and distance measures for binary data and their application in cluster analysis. There are 66 measures for binary data analyzed in the paper in order to provide a comprehensive insight into the problematics and to create their well-arranged overview. For this purpose, formulas by which they were defined are studied. In the next part of the research, the results of object clustering on generated datasets are compared, and the ability of measures to create similar or identical clustering solutions is evaluated. This is done by using chosen internal and external evaluation criteria, and comparing the assignments of objects into clusters in the process of hierarchical clustering. The paper shows which similarity measures and distance measures for binary data lead to similar or even identical results in hierarchical cluster analysis.

1 Introduction

A binary vector is one of the most common representations of patterns in datasets. Therefore many similarity and distance measures for binary data have been proposed over the years. They are usually well examined; they are often implemented in both – commercial (e.g., SPSS, MatLab) and non-commercial (e.g., R, Python) software. This paper focuses on associations among selected similarity measures and distance measures for binary data in the field of cluster analysis. Similarity measures and distance

[1]Department of Statistics and Probability, Faculty of Informatics and Statistics, University of Economics, Prague, Czech Republic; jana.cibulkova@vse.cz

[2]Department of Statistics and Probability, Faculty of Informatics and Statistics, University of Economics, Prague, Czech Republic; zdenek.sulc@vse.cz

[3]Department of Statistics and Probability, Faculty of Informatics and Statistics, University of Economics, Prague, Czech Republic; hana.rezankova@vse.cz

[4]Department of Statistics and Probability, Faculty of Informatics and Statistics, University of Economics, Prague, Czech Republic; xsirs00@vse.cz

measures for binary data are routinely used not only in the cluster analysis of binary data but in clustering nominal data as well. In fact, it is considered as standard procedure to perform a binary transformation when clustering nominal data.

The demand for processing binary (or nominal) data caused that numerous similarity measures and distance measures for binary data have been proposed over the years in various fields for various purposes. For example, the Jaccard similarity measure was designed for clustering flowers (Jaccard, 1901), while Forbes (1925) proposed a coefficient for clustering ecologically related species of fishes. However, an enormous quantity of available measures for binary data might be counterproductive, and it can cause confussion. Duplicities can easily occur, there might be a functional relationship between many measures for binary data, a certain measure can be referred to by several different names in the literature, etc.

On the one hand, there are many authors who defined their own measures for binary data, often by adjusting existing ones for the need of a specific study or not knowing that the measure already had been defined before (maybe in a different field of study). On the other hand, only a handful of authors studied and compared these measures among each other. Jackson et al. (1989) compared eight binary similarity measures before choosing the best measure for his ecological study. Hubálek (1982) collected and studied 43 measures. Twenty of them were used for cluster analysis on real data to produce five clusters of related measures. His study proved that many measures might lead to similar clustering solutions. Todeschini (2012) studied similarity measures for binary data on chemoinformatics dataset. Based on his study, he claims that binary data measures are often linearly dependent, and thus, the majority of them produces the same clusters. The best summary of similarity measures and distance measures for binary data so far was created by Cho and Chai (2010). They collected 76 measures used over the last century and revealed their relationships through a simple hierarchical clustering of values that were calculated by applying given measures on the same dataset. All of these authors provide some kind of survey of similarity measures; however, the studies contained a limited number of similarity measures or they were applied on one specific dataset. Furthermore, no one (except Hubálek, 1982) studied the measures formulas that create these associations in the first place.

This paper uses the studies of the previously mentioned authors as a baseline for further research that tries to be robust enough and based on mathematics and definitions that form these associations among measures for binary data. The paper aims to determine which measures for binary data lead to different clustering results and which ones lead to similar or even identical results of cluster analysis. Firstly, the theory of generating datasets, clustering method and evaluation criteria are briefly presented in Section 2. Also similarity and distance measures for binary data are introduced in this section. Section 3 is focused on revealing associations among the measures in their formulas. We expect the formulas that are functions of the same statistics to lead to the same or

similar results. This expectation is verified and analyzed later in the paper on clustering solutions of generated datasets. Measures are used in hierarchical clustering of generated datasets in Section 4. The results of object clustering on generated datasets are compared, the quality of clustering solutions is evaluated, and the ability of measures to create similar or identical clusters is examined. This approach makes our research different from previous research mentioned above.

2 Theoretical background

Theoretical background for the experiment, such as data generating process, similarity and distance measures for binary data, a chosen clustering method, and evaluation criteria are briefly introduced in this section. The vocabulary presented in the section is used in the rest of the paper.

2.1 Data generator

Numerous datasets with a given specific set of features are needed in order to make the results of the analysis reliable. For this purpose, the data generator suitable for the needs of the experiment is used (Cibulková and Řezanková, 2018). The data generator was designed mainly for generating multivariate datasets appropriate for cluster analysis.

A generated datasets consists of a given number of clusters, where each cluster corresponds to one sample of a given multivariate distribution. The idea of a dataset being a mixture of several samples from given multivariate distributions follows the logic of finite mixture models from model-based clustering. Finite mixture models assume that the population is made up of several distinct clusters, each following a different multivariate probability density distribution (Stahl and Sallis, 2012). Therefore, the problem of generating datasets with given features can actually be reduced to generating samples from given multivariate distributions.

The algorithm for generating samples from given multivariate distributions is inspired by the NORTA (NORmal-To-Anything) transformation in combination with Cholesky's decomposition. The NORTA algorithm, inspired by Cario and Nelson (1997), is used in order to generate samples from a given multivariate distribution. The Cholesky matrix transforms a vector of uncorrelated normally-distributed random variables into a vector of correlated normally-distributed random variables.

1. Generate a multivariate vector of uncorrelated normally-distributed random variables $Z^{ind} = (Z_1^{ind}, Z_2^{ind}, \ldots, Z_n^{ind})$.

2. Suppose, that $(\rho_{ij})_{i,j=1}^n$ is given correlation matrix. Positive definiteness of the matrix is verified and (if necessary) the matrix is adjusted to be the closest positive definite correlation matrix $(\rho_{ij}^*)_{i,j=1}^n$.

3. Get a multivariate standard-normal random vector $Z = (Z_1, Z_2, \ldots, Z_n)$, such that $Corr(Z_i, Z_j) = \rho^*_{ij}$, for $1 \leq i, j \leq n$ using Cholesky's decomposition following the algorithm described in (Higham, 2009).

4. Compute $U_i = \Phi(Z_i)$ for $i = 1, 2, \ldots, n$, where $\Phi(\cdot)$ is standard-normal cumulative distribution function.

5. Compute $X_i = F_i^{-1}(U_i)$ for $i = 1, 2, \ldots, n$, where F_i^{-1} is the inverse of given marginal cumulative distribution functions F_i.

Assuming each cluster in the dataset is generated from a given multivariate distribution, the generated dataset is a mixture of several samples obtained by this approach.

This generator allows us to generate numerous datasets with desired features to cover a wide range of datasets types, making the results of the analysis more robust.

2.2 Hierarchical clustering method and evaluation criteria

Agglomerative hierarchical cluster analysis is used in this paper. Its algorithm considers each object to start in its own cluster, and at each step, the nearest two clusters are combined into a higher-level cluster (Sokal and Michener, 1958).

The average linkage method was chosen for the experiment since it can be seen as a compromise between the sensitivity to outliers of complete linkage method and the tendency to form long chains (that do not correspond to the intuitive notion of clusters as compact, spherical objects) of single linkage method. This method takes average pairwise dissimilarity between objects in two different clusters. Let us denote $D_{average}\left(C_p, C_q\right)$ the distance between clusters C_p and C_q, with the number of objects n_p in the p-th cluster and n_q in the q-th cluster. Then dissimilarity between two clusters follows the formula:

$$D_{average}\left(C_p, C_q\right) = \frac{\sum_{\mathbf{x}_i \in C_p} \sum_{\mathbf{x}_j \in C_q} D\left(\mathbf{x}_i, \mathbf{x}_j\right)}{n_p n_q}, \tag{2.1}$$

where $D\left(\mathbf{x}_i, \mathbf{x}_j\right)$ is a distance between objects $\mathbf{x}_i$ and $\mathbf{x}_j$.

Since the analyses are performed on the generated datasets, and thus, the objects' cluster memberships are known, the produced clusters can be evaluated using both internal and external evaluation criteria.

The purity is an external evaluation criterion of cluster quality. It is the percentage of the total number of objects that were classified correctly. The larger the purity, the better the clustering performance. Let us suppose there are l classes (real clusters), while the clustering method generates k clusters, then the purity of the clustering solution with respect to the known classes is given by the formula:

$$purity = \frac{1}{n} \sum_{q=1}^{k} max\left(n_q^p\right), \tag{2.2}$$

where n is the total number of objects; n_q^p is the number of objects in the q-th cluster, that belongs to the p-th class; $(p = 1, 2, \ldots, l)$.

The entropy is an another external evaluation criterion. It is a measure of uncertainty, and the smaller the entropy, the better the clustering performance (Shannon, 1948). The entropy of the clustering solution with respect to the known classes follows the formula:

$$entropy = \frac{1}{n \ \log_2 l} \sum_{q=1}^{k} \sum_{p=1}^{l} n_q^p \log_2 \frac{n_q^p}{n_q}, \tag{2.3}$$

where n_q is the total number of objects in the q-th cluster $(q = 1, 2, \ldots, k)$.

The *Dunn index* is an internal evaluation criterion (Dunn, 1974). It is the ratio of the smallest distance between observations not in the same cluster to the largest intra-cluster distance. The Dunn index takes on values between zero and infinity and should be maximized. Let us denote a particular clustering partition $\mathcal{C} = C_1, C_2, \ldots, C_k$ of n objects into k disjoint clusters. Then the Dunn index is computed as:

$$Dunn\ index\ = \frac{\min_{C_q, C_p \in \mathcal{C}; C_q \neq C_p} \left(\min_{\mathbf{x}_i \in C_q; \mathbf{x}_j \in C_p} D(\mathbf{x}_i, \mathbf{x}_j) \right)}{\max_{C_q \in \mathcal{C}} diam\,(C_q)}, \tag{2.4}$$

where $diam(C_q)$ is the maximum distance between observations in cluster C_q.

The silhouette coefficient is an internal evaluation criterion which is calculated as the average of each object's silhouette width (which is usually displayed as the width in the silhouette graph). The silhouette width measures the degree of confidence in the clustering assignment of a particular object (Rousseeuw, 1987). The well-clustered objects do have values near 1 and poorly clustered objects have values near -1. For object x_i, it is defined as:

$$S\,(i) = \frac{b_i - a_i}{\max(b_i, a_i)}, \tag{2.5}$$

where a_i is the average distance between object x_i and all other objects in the same cluster, and b_i is the average distance between x_i and the objects in the nearest neighboring cluster following the formula $b_i = \min_{C_k \in \mathcal{C}; C_k \neq C(i)} \frac{\sum_{j \in C_k} D(\mathbf{x}_i, \mathbf{x}_j)}{n_k}$, where $C(i)$ is the cluster containing observation $\mathbf{x}_i$; $D(\mathbf{x}_i, \mathbf{x}_j)$ is the distance between observations $\mathbf{x}_i$ and $\mathbf{x}_j$; n_k number of objects in cluster C_k. The average silhouette width thus lies in the interval $[-1, 1]$, and should be maximized.

2.3 Similarity and distance measures for binary data

Let us denote the data matrix $\mathbf{X} = [x_{ic}]$, where $i = 1, 2, \ldots, n$ and $c = 1, 2, \ldots, m$; n is the total number of objects; m is the total number of variables. Suppose that two

objects, $\mathbf{x}_i$ and $\mathbf{x}_j$ are represented by the binary vector form. The symbols used for the numbers of variables with certain combinations of categories for objects, as shown in Table 1, are used for definitions of binary distance measures in this paper (Dunn and Everitt, 1982). In Table 1, a is the number of variables where the values of $\mathbf{x}_i$ and $\mathbf{x}_j$ are both equal to 1, meaning "positive matches", b is the number of variables where the value of $\mathbf{x}_i$ and $\mathbf{x}_j$ is $(0, 1)$, meaning "$\mathbf{x}_i$ absence mismatches", c is the number of variables where the value of $\mathbf{x}_i$ and $\mathbf{x}_j$ is $(1, 0)$, meaning "$\mathbf{x}_j$ absence mismatches", and d is the number of variables where both $\mathbf{x}_i$ and $\mathbf{x}_j$ are equal to 0, meaning "negative matches". Then, $m = a + b + c + d$.

Table 1: Symbols used for the numbers of variables with certain combinations of categories for objects $\mathbf{x}_i$ and $\mathbf{x}_j$

$\mathbf{x}_i \setminus \mathbf{x}_j$	1 (Presence)	0 (Absence)
1 (Presence)	a	b
0 (Absence)	c	d

Tables 2–4 provide the overview of formulas of the 66 similarity measures or distance measures for binary data. The main source for the formulas was Cho and Chai (2010). The second column in Tables 2–4 gives a definition of a measure using symbols from Table 1. The third column determines whether a measure is defined as a similarity measure or a distance measure.

The transformation from a similarity measure $S(\mathbf{x}_i, \mathbf{x}_j)$ into a distance measure $D(\mathbf{x}_i, \mathbf{x}_j)$, that is necessary in order to be able to calculate a proximity matrix in clustering process, is inspired by Deza and Deza (2013).

- If a similarity measure gets values from the interval $[0, 1]$, then the corresponding distance measure is obtained by the transformation $1 - S(\mathbf{x}_i, \mathbf{x}_j)$.

- If a similarity measure gets values from the interval $[-1, 1]$, then the corresponding distance measure is obtained by the transformation $\frac{1-S(\mathbf{x}_i,\mathbf{x}_j)}{2}$.

- If a similarity measure gets values from a different interval, let's say it is the interval $[min, max]$, then the similarity measure is transformed by min-max normalization in the first step, resulting that the values are from the interval $[0, 1]$. Then the corresponding distance measure is obtained by substracting 1 from the min-max normalized similarity measure.

Table 2: Similarity/distance measures for binary data overview – part 1

Measure	Formula	Type
JACCARD	$\frac{a}{a+b+c}$	similarity
DICE	$\frac{2a}{2a+b+c}$	similarity
CZEKANOWSKI	$\frac{2a}{2a+b+c}$	similarity
JACCARD_3W	$\frac{3a}{3a+b+c}$	similarity
NEI_LI	$\frac{2a}{(a+b)+(a+c)}$	similarity
SOKAL_SNEATH_I	$\frac{a}{a+2b+2c}$	similarity
SOKAL_MICHENER	$\frac{a+d}{a+b+c+d}$	similarity
SOKAL_SNEATH_II	$\frac{2a+2d}{2a+b+c+2d}$	similarity
ROGER_TANIMOTO	$\frac{a+d}{a+2(b+c)+d}$	similarity
FAITH	$\frac{a+0.5d}{a+b+c+d}$	similarity
GOWER_LEGENDRE	$\frac{a+d}{a+0.5(b+c)+d}$	similarity
INTERSECTION	a	similarity
INNERPRODUCT	$a+d$	similarity
RUSSEL_RAO	$\frac{a}{a+b+c+d}$	similarity
HAMMING	$b+c$	distance
EUCLID	$\sqrt{b+c}$	distance
SQUARED_EUCLID	$\sqrt{(b+c)^2}$	distance
CANBERRA	$(b+c)^{2/2}$	distance
MANHATTAN	$b+c$	distance
MEAN_MANHATTAN	$\frac{b+c}{a+b+c+d}$	distance
CITY_BLOCK	$b+c$	distance

Table 3: Similarity/distance measures for binary data overview – part 2

Measure	Formula	Type
MINKOWSKI	$(b+c)^{1/1}$	distance
VARI	$\frac{b+c}{4(a+b+c+d)}$	distance
SIZE_DIFFERENCE	$\frac{(b+c)^2}{(a+b+c+d)^2}$	distance
SHAPE_DIFFERENCE	$\frac{m(b+c)-(b-c)^2}{(a+b+c+d)^2}$	distance
PATTERN_DIFFERENCE	$\frac{4bc}{(a+b+c+d)^2}$	distance
LANCE_WILLIAMS	$\frac{b+c}{2a+b+c}$	distance
BRAY_CURTIS	$\frac{b+c}{2a+b+c}$	distance
HELLINGER	$2\sqrt{1-\frac{a}{\sqrt{(a+b)(a+c)}}}$	distance
CHORD	$\sqrt{2\left(1-\frac{a}{\sqrt{(a+b)(a+c)}}\right)}$	distance
COSINE	$\frac{a}{(\sqrt{(a+b)(a+c)})^2}$	similarity
GILBERT_WELLS	$\log(a)-\log(m)-\log\left(\frac{a+b}{m}\right)-\log\left(\frac{a+c}{m}\right)$	similarity
OCHIAI_I	$\frac{a}{\sqrt{(a+b)(a+c)}}$	similarity
FORBESI	$\frac{ma}{(a+b)(a+c)}$	similarity
FOSSUM	$\frac{m(a-0.5)^2}{(a+b)(a+c)}$	similarity
SORGENFREI	$\frac{a^2}{(a+b)(a+c)}$	similarity
MOUNTFORD	$\frac{a}{0.5(ab+ac)+bc}$	similarity
OTSUKA	$\frac{a}{((a+b)(a+c))^{0.5}}$	similarity
MCCONNAUGHEY	$\frac{a^2-bc}{(a+b)(a+c)}$	similarity
TARWID	$\frac{ma-(a+b)(a+c)}{ma+(a+b)(a+c)}$	similarity
KULCZYNSKI_II	$\frac{\frac{a}{2}(2a+b+c)}{(a+b)(a+c)}$	similarity
DRIVER_KROEBER	$\frac{a}{2}\left(\frac{1}{a+b}+\frac{1}{a+c}\right)$	similarity
JOHNSON	$\frac{a}{a+b}+\frac{a}{a+c}$	similarity
DENNIS	$\frac{ad-bc}{\sqrt{m(a+b)(a+c)}}$	similarity
SIMPSON	$\frac{a}{\min(a+b,a+c)}$	similarity
BRAUN_BANQUET	$\frac{a}{\max(a+b,a+c)}$	similarity
FAGER_MCGOWAN	$\frac{a}{\sqrt{(a+b)(a+c)}}-\frac{\max(a+b,a+c)}{2}$	similarity
FORBES_II	$\frac{ma-(a+b)(a+c)}{m\min(a+b,a+c)-(a+b)(a+c)}$	similarity

Table 4: Similarity/distance measures for binary data overview – part 3

Measure	Formula	Type
SOKAL_SNEATH_IV	$\frac{1}{4}\left(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{b+d}+\frac{d}{c+d}\right)$	similarity
GOWER	$\frac{a+d}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	similarity
PEARSON_I	χ^2; where $\chi^2=\frac{m(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$	similarity
PEARSON_II	$\left(\frac{\chi^2}{m+\chi^2}\right)^{\frac{1}{2}}$	similarity
PEARSON_III	$\left(\frac{\rho}{m+\rho}\right)^{\frac{1}{2}}$; where $\rho=\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	similarity
PEARSON_HERON_I	$\rho=\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	similarity
PEARSON_HERON_II	$\cos\left(\frac{\pi\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}\right)$	similarity
SOKAL_SNEATH_III	$\frac{a+d}{b+c}$	similarity
SOKAL_SNEATH_V	$\frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	similarity
COLE	$\frac{\sqrt{2}(ad-bc)}{\sqrt{(ad-bc)^2-(a+b)(a+c)(b+d)(c+d)}}$	similarity
STILES	$\log_1 0\frac{m\left(\lvert ad-bc\rvert-\frac{n}{2}\right)^2}{(a+b)(a+c)(b+d)(c+d)}$	similarity
OCHIAI_II	$\frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	similarity
YULE_Q	$\frac{2bc}{ad+bc}$	distance
YULE_W	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}-\sqrt{bc}}$	similarity
KULCZYNSKI_I	$\frac{a}{b+c}$	similarity
TANIMOTO	$\frac{a}{(a+b)+(a+c)-a}$	similarity
HAMANN	$\frac{(a+d)-(b+c)}{a+b+c+d}$	similarity
MICHAEL	$\frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$	similarity

3 Associations in formulas

It is possible to detect several groups of measures defined in Section 2.3 where functional relationships among measures exist. The formulas were studied and compared in order to unhide any functional relationships among them. Various associations were discovered when studying the formulas. These associations may provide the first categorization of similarity and distance measures for binary data. We may consider four groups of measures:

1. Euclid-based measures,
2. Pearson-based measures,
3. Hellinger-based measures,
4. other measures.

3.1 Euclid-based measures

Numerous measures for binary data proved to be dependent on the Euclid distance. Let us denote the squared Euclid distance measure as

$$eu = (b + c), \tag{3.1}$$

where b and c are symbols used for a number of variables with certain combinations of categories for objects $\mathbf{x}_i$ and $\mathbf{x}_j$ described in Table 1. Table 5 shows, how are some of measures, presented in Section 2.3, dependent on eu.

The first column contains a measure's name, a measure is expressed in relation to the squared Euclid distance measure in the second column. If there is any restriction for the relationship between a measure from the first column and the squared Euclid distance measure, this restriction would be noted in the third column.

3.2 Pearson-based measures

Other measures for binary data proved to be dependent on the Pearson correlation coefficient. Let's denote

$$\rho = \frac{ad - bc}{\sqrt{(a + b)(a + c)(b + d)(c + d)}}, \tag{3.2}$$

where a, b, c, d are symbols used for number of variables with certain combinations of categories for objects $\mathbf{x}_i$ and $\mathbf{x}_j$ described in Table 1. The measures that are based on the Pearson correlation coefficient can be re-written using ρ in their formulas as shown in Table 7. Measures' names are in the first column, measures are expressed in relation to ρ in the second column.

Table 5: Association among the measures based on Euclid distance

Measure	Formula	Condition
SQUARED_EUCLID	eu	-
EUCLID	$\sqrt{eu}$	-
HAMMING	eu	-
CANBERRA	eu	-
MANHATTAN	eu	-
MEAN_MANHATTAN	$\frac{eu}{m}$	-
CITY_BLOCK	eu	-
MINKOWSKI	eu	-
VARI	$\frac{eu}{4m}$	-
SOKAL_MICHENER	$1 - \frac{eu}{m}$	-
INNERPRODUCT	$m - eu$	-
SIZE_DIFFERENCE	$\frac{eu^2}{m^2}$	-
HAMMANN	$\frac{m-2eu}{m}$	-
GOWER_LEGENDRE	$\frac{2(m-eu)}{2m-eu}$	-
SOKAL_SNEATH_II	$\frac{2(m-eu)}{2m-eu}$	-
SOKAL_SNEATH_III	$\frac{n}{eu} - 1$	-
ROGER_TANIMOTO	$\frac{m-eu}{m+eu}$	-
FAITH	$\frac{m-eu-0.5d}{m}$	-
SHAPE_DIFFERENCE	$\frac{eu}{m}$	if $b = c$

Table 6: Association among the measures based on Pearson distance

Measure	Formula
PEARSON_HERON_I	ρ
PEARSON_I	$n\rho^2$
PEARSON_II	$\left(\frac{(n\rho^2)}{n+(n\rho^2)}\right)^{\frac{1}{2}}$
PEARSON_III	$\left(\frac{\rho}{n+\rho}\right)^{\frac{1}{2}}$

3.3 Hellinger-based measures

Some distance and similarity measures for binary data are dependent on the Hellinger distance. Formulas of these measures demonstrate the dependency in Table 7.

Table 7: Association among the measures based on Hellinger distance

Measure	Formula
HELLINGER	hel
CHORD	$\frac{hel}{\sqrt{2}}$
OCHIAI_I	$1-\left(\frac{hel}{2}\right)^2$
OTSUKA	$1-\left(\frac{hel}{2}\right)^2$
FAGER_MCGOWAN	$1-\left(\frac{hel}{2}\right)^2-\frac{\max(a+b,a+c)}{2}$
SORGENFREI	$\left[1-\left(\frac{hel}{2}\right)^2\right]$
FORBESI	$\frac{m}{a}\left[1-\left(\frac{hel}{2}\right)^2\right]$

Measures' names are in the first column, measures are expressed in relation to the Hellinger distance in the second column. Let us denote

$$hel = 2\sqrt{1-\frac{a}{\sqrt{(a+b)(a+c)}}}, \tag{3.3}$$

where a, b, c and d are symbols used for number of variables with certain combinations of categories for objects $\mathbf{x}_i$ and $\mathbf{x}_j$ described in Table 1.

3.4 Other measures and associations among them

Several groups up to four measures of similarity or distance measures from Section 2.3 seem to have some kind of dependency among themselves.

The JACCARD similarity measure is identical with similarity measure TANIMOTO.

$$\begin{aligned} \text{JACCARD} &= \frac{a}{a+b+c} = \\ &= \text{TANIMOTO} = \frac{a}{(a+b)+(a+c)-a} \end{aligned} \tag{3.4}$$

Similarity measures DICE, CZEKANOWSKI, NEI_LI are identical.

$$\begin{aligned} \text{DICE} &= \frac{2a}{2a+b+c} = \\ &= \text{CZEKANOWSKI} = \frac{2a}{2a+b+c} = \\ &= \text{NEI\_LI} = \frac{2a}{(a+b)+(a+c)} \end{aligned} \tag{3.5}$$

Another pair of similarity measures that are defined by the same formula is LANCE_WILLIAMS and BRAY_CURTIS.

$$\begin{aligned} \text{LANCE\_WILLIAMS} &= \frac{b+c}{2a+b+c} = \\ &= \text{BRAY\_CURTIS} = \frac{b+c}{2a+b+c} \end{aligned} \tag{3.6}$$

Moreover, if $2a = b + c$, then all five measures DICE, CZEKANOWSKI, NEI_LI, LANCE_WILLIAMS and BRAY_CURTIS would be equal to each other.

There is a linear dependency between similarity measures INTERSECTION and RUSSEL_RAO.

$$\begin{aligned} \text{INTERSECTION} &= a = \\ &= m\,(\text{RUSSEL\_RAO}) = m\left(\frac{a}{a+b+c+d}\right) \end{aligned} \tag{3.7}$$

Similarity measures KULCZYNSKI_II and DRIVER_KROEBER are identical and they are linearly dependent with JOHNSON similarity measure.

$$\begin{aligned} \text{KULCZYNSKI\_II} &= \frac{\frac{a}{2}(2a+b+c)}{(a+b)(a+c)} = \\ &= \text{DRIVER\_KROEBER} = \frac{a}{2}\left(\frac{1}{a+b}+\frac{1}{a+c}\right) = \\ &= 0.5\,(\text{JOHNSON}) = 0.5\left(\frac{a}{a+b}+\frac{a}{a+c}\right) \end{aligned} \tag{3.8}$$

Similarity measures SOKAL_SNEATH_V and OCHIAI_II are identical.

$$\begin{aligned} \text{SOKAL\_SNEATH\_V} &= \frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} = \\ = \text{ OCHIAI\_II} &= \frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \end{aligned} \tag{3.9}$$

If $d = 0$, then formulas of measures KULCZYNSKI_II and SOKAL_SNEATH_IV are identical.

$$\begin{aligned} \text{KULCZYNSKI\_II} &= \frac{\frac{a}{2}(2a+b+c)}{(a+b)(a+c)} = \\ = \text{SOKAL\_SNEATH\_IV } &= \frac{1}{4}\left(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{b+d}+\frac{d}{c+d}\right) \end{aligned} \tag{3.10}$$

If $d = 0$, then similarity measures KULCZYNSKI_I and SOKAL_SNEATH_III are identical.

$$\begin{aligned} \text{SOKAL\_SNEATH\_III} &= \frac{a+d}{b+c} = \\ = \text{KULCZYNSKI\_I} &= \frac{a}{b+c} \end{aligned} \tag{3.11}$$

4 Associations in clustering solutions

We have shown that some measures are functionally dependent, linearly dependent, or even identical in Section 3. Therefore, we may expect that the process of assigning objects into clusters would be somehow associated. If there is a linear dependence between measures, we may even assume that the clustering process will be identical. Thus, it would lead to the same clustering solutions and that the dendrograms will be similar. Figure 1 demonstrates nearly identical outcome of hierarchical clustering process (average-linkage) on a sample dataset when four different (but linearly dependent) similarity measures were used. The goal of this section is to express such a similarity of dendrograms for various measures in numerical terms.

4.1 Experiment design

The datasets were generated using the data generator described in Section 2.1. Desired features of the generated datasets were chosen with an aim to cover a wide range of possible situations that can occur. Since the paper focuses on measures suitable for binary data, all data are generated from the Bernoulli distribution, but they differ in a number of objects, a number of variables, and the strength of a relationship among

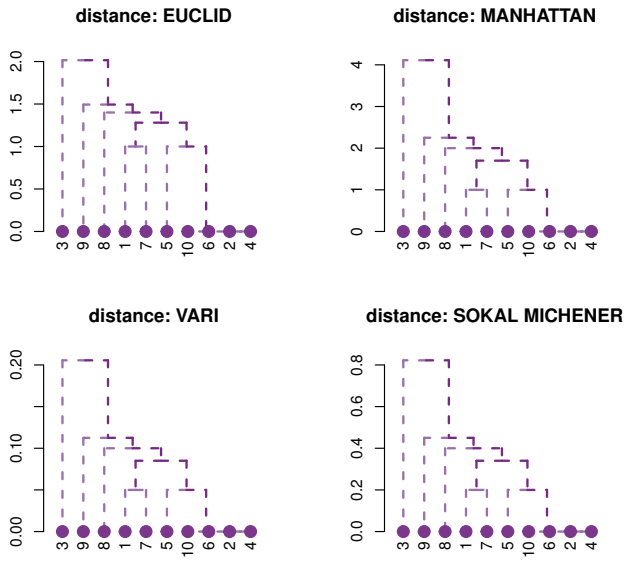

Figure 1: Example of 4 different measures used in hierarchical clustering of a sample dataset leading to the same result

variables (expressed by correlation matrix). The following features were used in the data-generating process:

1. probability distribution: **Bernoulli**,
2. number of variables: **5, 10, 15, 20**,
3. number of clusters: **2, 3, 4, 5**,
4. number of observations in a cluster: **30–60, 60–120, 30-120**,
5. correlation among variables: $\in [-1; 1]$**; at random for each cluster**.

This leads to 48 possible combinations of dataset features (numbers of variables, numbers of clusters, and numbers of observations). Each combination of features is generated ten times to make the outcomes of the experiment more robust, and the final number of generated datasets is hence equal to 480. The information about "real cluster" membership is stored for each observation of each dataset, and it is later used in the process of evaluation of clustering solutions.

The average linkage method combined with every binary similarity distance from Section 2.3 is applied to each dataset. Clustering solutions with three clusters are used to examine associations among similarity and distance measures in clustering solutions via internal evaluation criteria (average silhouette width, Dunn index), in order to mimic a real-life situation, where the number of real clusters is usually unknown. The examination of associations among measures in clustering solutions via external evaluation

criteria (purity, entropy) uses information about "real cluster" membership. These values of four evaluation criteria from Section 2.2 are calculated for each clustering solution, and their averages and standard deviations across all datasets are compared with respect to similarity/distance measure used.

Due to the complexity of computations, the evaluation of the assignment of objects into clusters is performed on two-to-four-clusters clustering solutions. For each number of clusters in the clustering solution (the last three steps of the clustering process) and each pair of measures from Section 2.3, the percentage of identical objects-into-clusters assignments was calculated. The percentage gives a value between 0 and 1, where 1 means the two clustering outcomes match identically. The percentage of identical objects-into-clusters assignments is also known as the *Rand index* (Rand, 1971). The pairs of last three steps of the clustering process for every two similarity measures are then compared, and measures that lead to identical/similar clustering process are identified.

All the calculations and visualizations were done in R programming language (R Core Team, 2018).

4.2 Results of the experiment

There are 66 similarity and distance measures for binary data examined in the paper and used in the clustering process. The ability of the measures to produce a quality clustering solution is measured by four chosen evaluation criteria: purity, entropy, Dunn index, and silhouette coefficient.

Average values of evaluation criteria (calculated as described in Section 4.1) are shown in Figure 2. In general, Euclid-based measures (red color in all graphs) produce clusters with the highest quality according to the evaluation criteria. The quality of Pearson-based measures (purple color in all graphs) is, on average, not very good, but it is steady across all four measures. The same applies to Hellinger-based measures (black color in all graphs), and all pairs/groups of measures, where any functional relationship was detected in Section 3. Alhough some measures lead to clustering solutions with low average quality, keep in mind that these measures might perform very well when clustering datasets with specific features. Therefore, one should not jump into any conclusions and general recommendations for universal use.

Even though the evaluation criteria' average values might not be beneficial for formulating general recommendations, they help reveal associations among the measures. The relationships among binary similarity/distance measures are quite evident based on the average values of evaluation criteria and their standard deviations. Figure 3 shows the arrangement of the clusters produced by average-linkage hierarchical clustering with Euclid distance in the dendrogram. From this graph, it is possible to see that some measures lead to more similar clustering solutions than others. All Euclid-based measures lead to clustering solution with similar values of evaluation criteria, except for

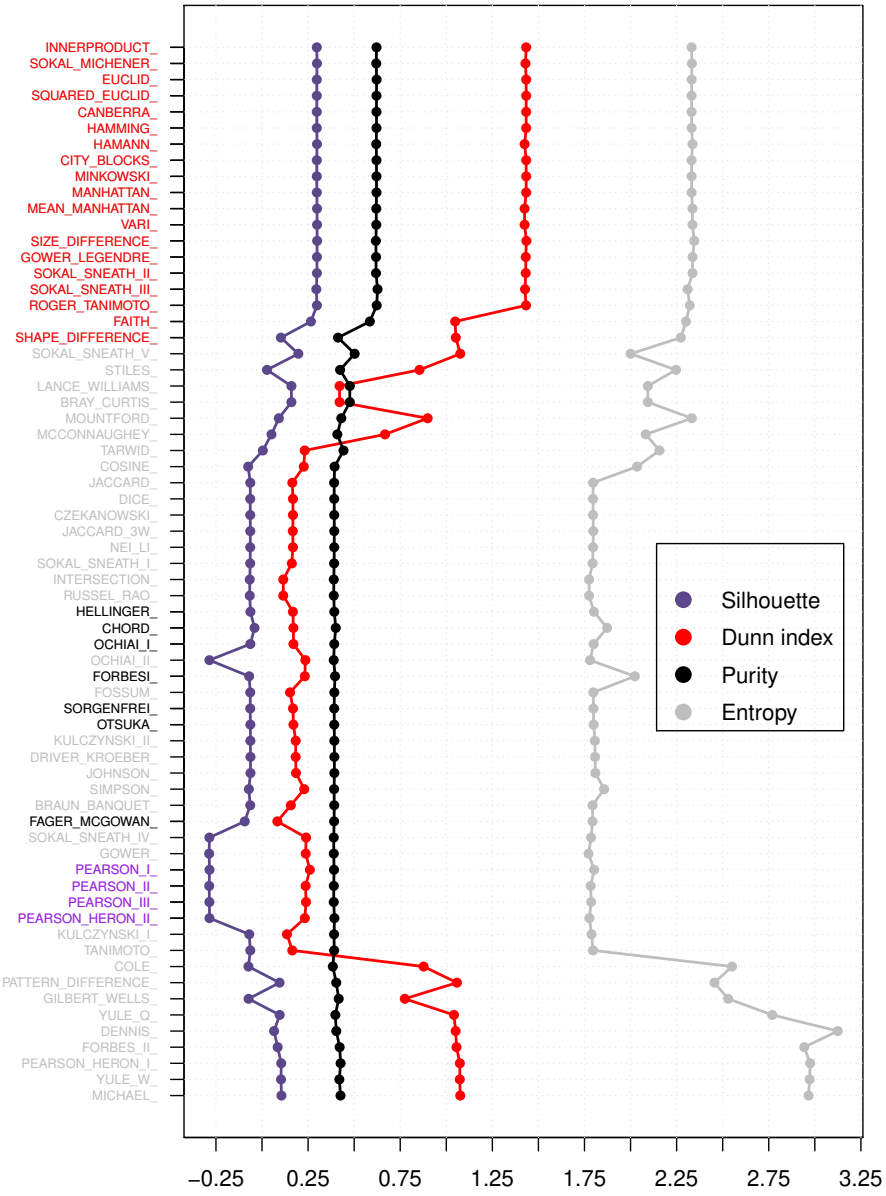

Figure 2: Average value of evaluation criteria representing clusters quality of examined similarity or distance measures for binary data

SHAPE_DIFFERENCE. That can be explained by the fact, that there exists a functional relationship among all these measures, but only SHAPE_DIFFERENCE has this relationship conditiond by a restriction (see Section 3.1). All identical or linearly dependent similarity and distance measures, see Section 3, lead to the identical cluster solutions as well. Clearly, the Euclid-based measures lead to very similar clustering solutions in terms of cluster quality. Due to non-linear relationships among Pearson-based measures, all the Pearson-based measures are much less similar in clustering solution quality. Negative-matches exclusive measures (such as JACCARD, HELLINGER, DICE, ...) also produce clusters of similar quality. Measures for which we have not found any association to other measures generally provide qualitatively different clustering solutions.

Figure 4 shows the average percentages of identical objects-into-clusters assignments for every pair of similarity or distance measures for binary data. The figure reflects the average similarity of dendrograms in the final three steps of the clustering process (2–4 clusters) for any two measures. The associations among all measures in the clustering process are indisputable. Almost all measures lead to more or less similar clustering solutions (at least in the last three steps of the clustering process). Especially dendrograms for identical or linearly dependent measures are essentially all the same. At the same time, it can be seen that measures without any apparent dependency on any other measure (such as DENNIS, MICHAEL, or MOUNTFORD) produce least similar dendrograms. Interestingly, the absence of negative matches in a measure's formula and the presence of only positive matches in a measure's formula seems to be a common feature for measures that produced clustering solutions of similar quality and almost identical dendrograms. These negative-matches exclusive measures produce very similar dendrograms regardless of whether they are Pearson-based or Hellinger-based.

5 Discussion and conclusion

Many similarity measures and distance measures for binary data have been used in various fields of study. This led to a situation where measures are duplicated or strongly dependent on each other. Despite the fact that these measures are widely known and often used, only a limited number of authors aimed their research in this area. Even though their studies usually focused on a limited number of similarity/distance measures for binary data or they were applied on only one specific dataset, they concluded that binary data measures are often linearly dependent, and thus, they often produce the same clusters.

Not only were we able to support this claim on hundreds of generated datasets, but we also explained the reason for this behavior (that is rooted in the measures' definition) while providing a comprehensive study of 66 selected measures for binary data. Measures were examined based on the quality of clustering solutions they produce and based

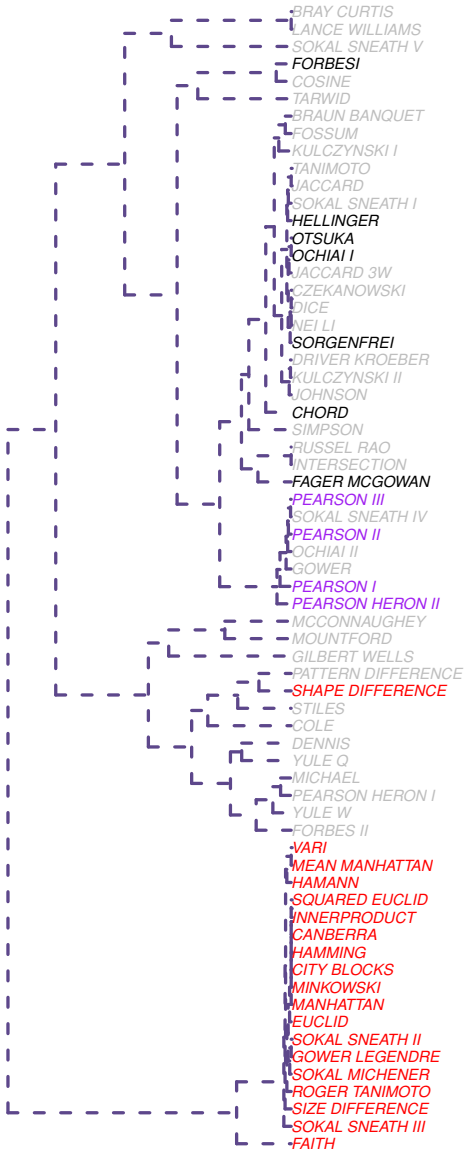

Figure 3: Dendrogram representing clusters quality of examined similarity or distance measures for binary data

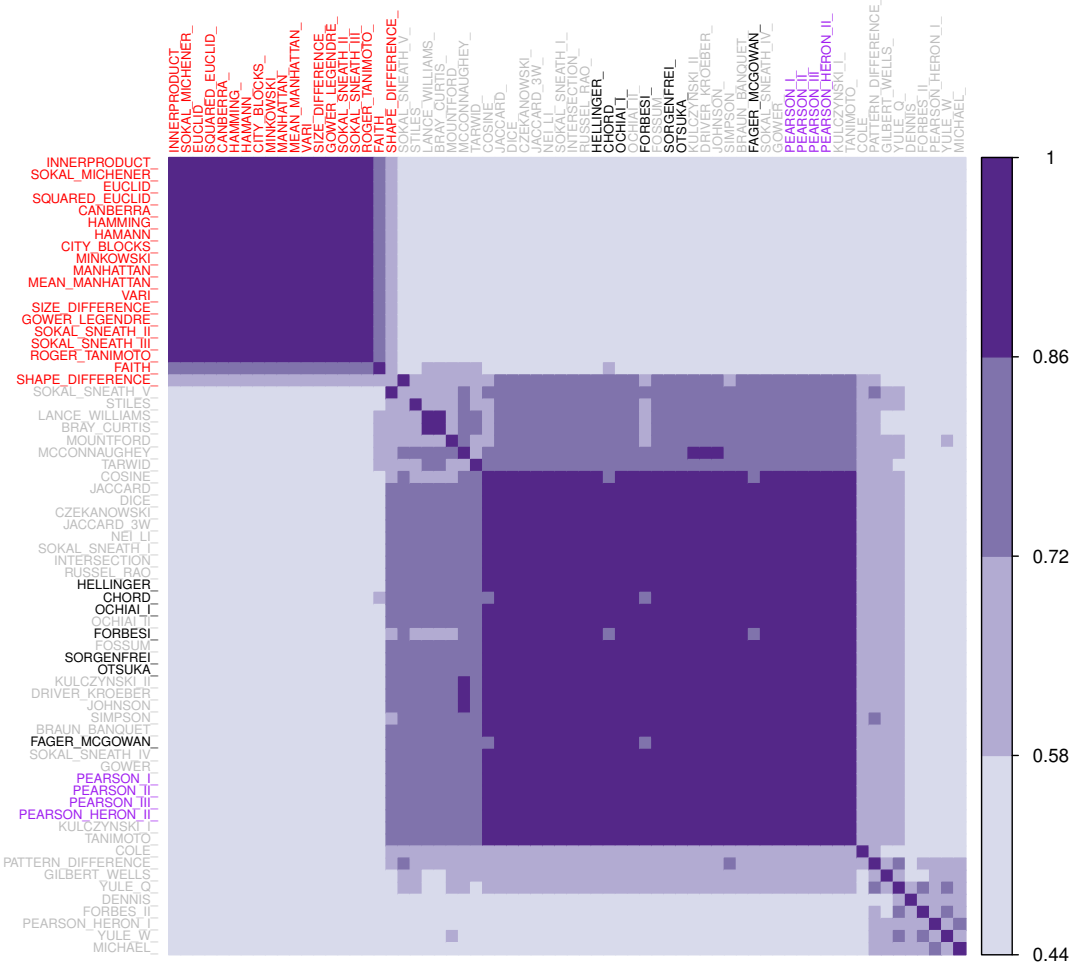

Figure 4: The average percentage of identical objects-into-clusters assignations

on the percentage of identical objects-into-clusters assignations in the last three steps of the hierarchical clustering process. Based on this we might claim that Euclid-based measures lead to identical (or very similar) clustering solutions. Another big group of measures that lead to very similar clustering results are negative-matches exclusive measures.

Acknowledgment

This work was supported by the University of Economics, Prague under Grant IGA F4/44/2018.

References

[1] Cairo, M. and Nelson, B. (1997): *Modeling and Generating Random Vectors with Arbitrary Marginal Distributions and Correlation Matrix*. Technical Report. Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL.

[2] Charu, C.A. and Chandan, K.R. (2013): *Data clustering: Algorithms and applications*. Boca Raton, FL: Chapman & Hall.

[3] Chen, H. (2001): Initialization for NORTA: Generation of random vectors with specified marginals and correlations. *INFORMS Journal on Computing*, **13**(4), 312–331.

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

[5] Cibulková, J. and Řezanková, H. (2018): Categorical Data Generator. In T. Loster, T. Pavelka (eds.): *International Days of Statistics and Economics 2018*. Slaný: Melandrium, 288–296.

[6] Deza, M.M. and Deza, E. (2013): *Encyclopedia of Distances*. New York, NY: Springer.

[7] Dunn, G. and Everitt, B.S. (1982): *An Introduction to Mathematical Taxonomy*. Cambridge: Cambridge University Press.

[8] Forbes, S. A. (1925): Method of determining and measuring the associative relations of species. *Science*, **61**, 5–24.

[9] Higham, N.J. (2009): Cholesky factorization. *Wiley Interdisciplinary Reviews: Computational Statistics*, **1**(2), 251–254.

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

[11] Rand, W.M. (1971): Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**, 846–850.

[12] Rousseeuw P.J. (1987): Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 53–65.

[13] Shannon, C.E. (1948): A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379–423.

[14] Jaccard, P. (1901): Etude comparative de la distribuition florale dans une portion des Alpes et des Jura. *Bulletin de la Société vaudoise des sciences naturelles*, **37**, 547–579.

[15] Jackson, D.A., Somers, K.M. and Harvey, H.H. (1989): Similarity coefficients: Measures of co-occurrence and association or simply measures of occurrence?. *The American Naturalist*, **133**(3), 436–453.

[16] R Core Team (2018): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`.

[17] Sokal R. and Michener, C., (1958): A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, **38**, 1409–1438.

[18] Stahl, D. and Sallis, H. (2012): Model-based cluster analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, **4**(4), 341–358.

[19] Todeschini, R., Consonni, V., Xiang, H., Holliday, J., Buscema, M. and Willet, P. (2012): Similarity coefficients for binary chemoinformatics Data: Overview and extended comparison using simulated and real data sets. *Journal of Chemical Information and Modeling*, **52**(11), 2884–2901.