Data Science: Similarity, Dissimilarity and Correlation Functions

Ildar Z. Batyrshin(✉)

Instituto Politécnico Nacional, Centro de Investigación en Computación,
Av. Juan de Dios Bátiz S/N, Nueva Industrial Vallejo,
07738 Ciudad de México, CDMX, Mexico
batyrl@gmail.com

Abstract. The lecture presents a new, non-statistical approach to the analysis and construction of similarity, dissimilarity and correlation measures. The measures are considered as functions defined on an underlying set and satisfying the given properties. Different functional structures, relationships between them and methods of their construction are discussed. Particular attention is paid to functions defined on sets with an involution operation, where the class of (strong) correlation functions is introduced. The general methods constructing new correlation functions from similarity and dissimilarity functions are considered. It is shown that the classical correlation and association coefficients (Pearson's, Spearman's, Kendall's, Yule's Q, Hamann) can be obtained as particular cases.

Keywords: Similarity measure · Pearson's product-moment correlation · Spearman's rank correlation · Kendall's rank correlation · Yule's Q

1 Introduction

Data Science is the buzzword of the last decade. One of the goals of Data Science is to extract knowledge, insights and usable information from various types of data. Data Science courses typically include the topics on statistical data analysis, machine learning methods and data mining. Correlation, association, similarity, relationship and interestingness coefficients or measures play an important role in data analysis, classification tasks and data mining. Dozens of such measures have been created for various types of data, and articles on these indicators are cited in hundreds and thousands of papers [3, 13, 14, 20, 23–25, 27]. Such measures are defined for dichotomous or real valued variables measured for sampling units, for rankings, binary or real valued vectors of attributes, time series, fuzzy sets of different types, probabilistic distributions, images, rating profiles etc. These measures used in information retrieval, data classification, machine learning, analysis of relationships and decision making in ecology, computational linguistics, image and signal processing, financial data analysis, bioinformatics and social sciences. Often, measures introduced for one type of data give misleading results when they are used for analysis of relationships between data of another type [8]. There is a need in construction of a general theory of similarity, correlation and association measures, which will be able to analyze general

G. S. Osipov et al. (Eds.): Artificial Intelligence, LNAI 11866, pp. 13–28, 2019.
https://doi.org/10.1007/978-3-030-33274-7_2

properties of these measures independent on domain or for large class of data types. Such a theory will make it possible to transfer the methods of constructing association measures from one domain to another and propose such measures for new data types.

The first steps in construction of the general theory of similarity, dissimilarity and correlation measures have been done in [7, 9, 10] where the fundamental results on definition and construction of association measures (correlation coefficients) on sets with involution operations have been proposed. It was shown that many known correlation coefficients can be considered as functions defined on an underlying set with involution operation and satisfying a simple set of properties. The general methods of construction of these functions from suitable similarity or dissimilarity functions have been proposed. These results can be considered as an alternative solution of the Kendall's problem of construction of general correlation coefficient [22]. Kendall proposed a formula that gives as particular cases Pearson's product-moment correlation, Spearman's and Kendall's rank correlation coefficients but it was not clear how to obtain from this formula other known association and correlation coefficients and how to build correlation coefficients for other domains. The approach proposed in [7, 9, 10] gives possibility to construct the classical correlation coefficients that can be obtained from Kendall's formula, the Yule's Q and Hamann coefficients together with some other association coefficients for binary data and gives a tool for constructing new correlation coefficients on new domains.

Similarity, dissimilarity and correlation measures or coefficients are considered in this lecture as functions defined on an underlying set Ω and satisfying some reasonable sets of properties. Generally, these functions can be considered as *association functions*, measuring (may be non-statistical) associations or relationships between objects. The methods of construction of such functions, the transformations of them and relationships between them studied in [7, 9, 10]. This paper presents a short and updated description of some part of author's Lecture on RAAI Summer School 2019. Due to the limit on the size of the paper it was not possible to include all topics discussed in [9] and in this lecture. The presented paper can be considered as complementary to the [9]. It includes new methods of construction of correlation functions presented recently on INES 2019 [10] and contains more examples of (dis)similarity and correlation functions illustrating these methods. The list of references includes only several works. Hundreds or thousands papers have been published on related topics during more than one hundred years. Of course it was not possible to include most of them. Some useful references can be found in the papers cited in the list of references.

2 Examples of Similarity Measures and Correlation Coefficients for Different Domains

2.1 Binary *n*-Tuples

Consider n-tuple $x = (x_1, \ldots, x_n)$, $x_i \in \{0, 1\}, i = 1, \ldots, n$. It can represent an object x described by n binary features, attributes or properties such that $x_i = 1$ if the object x possesses the i-th attribute and $x_i = 0$ in the opposite case. Denote X the set of attributes possessed by x. In statistics, a binary n-tuple can denote n measurements of a

dichotomous variable or property x for n sampling units. $x_i = 1$ denotes that the property x fulfilled for the unit i, and $x_i = 0$ otherwise. We use here the first interpretation of binary n-tuples. For two binary n-tuples $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$ denote:

- a the number of attributes possessed by both objects x and y, when $x_i = y_i = 1$;
- b the number of attributes possessed by x and not by y, when $x_i = 1, y_i = 0$;
- c the number of attributes possessed by y and not by x, when $x_i = 0, y_i = 1$;
- d the number of attributes not possessed by both objects x and y, when $x_i = y_i = 0$.

The numbers a and d usually referred to as the numbers of *positive and negative matches*, respectively. Below are examples of similarity and association measures for binary n-tuples [13]:

Jaccard similarity measure:

$$S_J = \frac{a}{a+b+c} = \frac{|X \cap Y|}{|X \cup Y|},$$

Simple Matching similarity measure:

$$S_{SM} = \frac{a+d}{a+b+c+d} = \frac{|X \cap Y| + |\bar{X} \cap \bar{Y}|}{n},$$

Hamann coefficient:

$$A_H = \frac{(a+d)-(b+c)}{a+b+c+d} = \frac{|X \cap Y| + |\bar{X} \cap \bar{Y}| - |X \cap \bar{Y}| - |\bar{X} \cap Y|}{n},$$

Yule's Q association coefficient:

$$A_{Y-Q} = \frac{ad-bc}{ad+bc} = \frac{|X \cap Y| \cdot |\bar{X} \cap \bar{Y}| - |X \cap \bar{Y}| \cdot |\bar{X} \cap Y|}{|X \cap Y| \cdot |\bar{X} \cap \bar{Y}| + |X \cap \bar{Y}| \cdot |\bar{X} \cap Y|}.$$

2.2 Real Valued n-Tuples

Consider n-tuples $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$ with real valued components.

Cosine similarity measure:

$$cos(x, y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}},$$

where $x_i, y_i \geq 0$, $i = 1, \ldots, n$.

Pearson's product-moment correlation coefficient:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

For both measures, it is supposed that denominators do not equal to zero.

2.3 Rankings

Consider two rankings of n objects $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$ containing n different integer ranks, $1 \le x_i, y_i \le n$, i.e. "without ties".

Spearman's rank correlation coefficient [12, 18, 22]:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

where $d_i = x_i - y_i, i = 1, \ldots, n$.

Kendall's rank correlation coefficient [12, 18, 22] for real valued n-tuples $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$ without ties, i.e. $x_i \neq x_j$, $y_i \neq y_j$ for all i, j:

$$\tau = \frac{NC - ND}{n(n-1)/2},$$

where NC is the *number of concordant pairs* (i, j) calculated as follows:

$$NC = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} s_{ij}, \quad \text{and} \quad s_{ij} = \begin{cases} 1, & if\ (x_i - x_j)(y_i - y_j) > 0 \\ 0, & otherwise \end{cases},$$

and ND is the *number of disconcordant pairs* calculated as follows:

$$ND = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_{ij}, \quad \text{and} \quad d_{ij} = \begin{cases} 1, & if\ (x_i - x_j)(y_i - y_j) < 0 \\ 0, & otherwise \end{cases}.$$

Note that only the ordering of the component values of n-tuples is used.

2.4 Finite Probabilistic Distributions

Suppose $x = (x_1, \ldots, x_n)$ is a finite probabilistic distribution, i.e. $x_i \ge 0$, for all $i = 1, \ldots, n$, and $\sum_{i=1}^{n} x_i = 1$.

Bhattacharyya coefficient is defined as follows [1]:

$$S(x,y) = \sum_{i=1}^{n} \sqrt{x_i y_i}.$$

2.5 Kendall's General Correlation Coefficient

Kendall [22] considered the problem of constructing *general correlation coefficient*. For two n-tuples $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ he defined it as:

$$r = \frac{\sum_{i,j=1}^{n} a_{ij} b_{ij}}{\sqrt{\sum_{i,j=1}^{n} a_{ij}^2 \sum_{i,j=1}^{n} b_{ij}^2}},$$

where the values a_{ij} calculated from the values x_i and x_j, similarly b_{ij} calculated from y_i and y_j, such that the following conditions are fulfilled: $a_{ij} = -a_{ji}$, $b_{ij} = -b_{ji}$, $a_{ii} = b_{ii} = 0$ for all $i, j = 1, \ldots, n$. As particular cases of this formula he obtained Spearman's and Kendall's rank correlation coefficients, and Pearson's product-moment correlation coefficient. How to obtain other correlation coefficients from this formula does not clear.

3 Similarity and Dissimilarity Functions

The examples of similarity measures and correlation coefficients considered in the previous section show that there is a variety of such indicators for different types of data. Generally, there are introduced tens of such measures and coefficients [13, 14, 27]. To analyze the general properties of these measures, possible relationships between them and to study the methods of their construction in [9] it was proposed to consider these measures as functions defined on a universal domain Ω and satisfying some simple sets of properties. Three main types of such functions have been considered: similarity, dissimilarity and correlation functions, which can be used as models of similarity, dissimilarity and correlation measures and coefficients, respectively. Below we introduce these functions and consider some of their properties. More results can be found in [9].

Further, for brevity, similarity and dissimilarity functions will be referred to as *(dis) similarity functions* or as *resemblance functions*.

3.1 Definition of Similarity, Dissimilarity and Correlation Functions

Let Ω be a nonempty set that will be referred to as a *universal domain* or an *underlying set* in definition of similarity and correlation functions. As Ω we can consider domain specific for a considered data type: the set of all binary n-tuples, the set of all real valued vectors of the length n, the set of all fuzzy sets defined on some domain X, the set of some images or objects considered in some problem, etc.

A function $S : \Omega \times \Omega \to [0, 1]$ is called a *similarity function* on Ω if for all x, y in Ω it is *symmetric*:

$$S(x, y) = S(y, x),$$

and *reflexive*:

$$S(x, x) = 1.$$

A function $D : \Omega \times \Omega \to [0, 1]$ is a *dissimilarity function* on Ω if for all x, y in Ω it is *symmetric*:

$$D(x, y) = D(y, x),$$

and *irreflexive*:

$$D(x, x) = 0.$$

These functions are called *complementary* if for all x, y in Ω it is fulfilled:

$$S(x, y) + D(x, y) = 1.$$

For complementary (dis)similarity functions we have:

$$S(x, y) = 1 - D(x, y), D(x, y) = 1 - S(x, y). \quad (1)$$

A function $A : \Omega \times \Omega \to [-1, 1]$ is a *correlation function* (association measure) on Ω if for all x, y in Ω it is *symmetric*:

$$A(x, y) = A(y, x),$$

reflexive:

$$A(x, x) = 1,$$

and negative: $A(x, y) < 0$ for some x, y in Ω.

Such correlation functions will be referred to as *weak* correlation functions if they will not satisfy inverse relationship property considered in Sect. 4. In Sect. 4 we define a strong (invertible) correlation function on a set Ω with involution operation.

It is easy to see that the indicators considered in the previous section belong to the following classes of functions:

- **Similarity functions:** Jaccard, Simple Matching, Cosine (if all x_i and y_i have non-negative values) similarity measures and Bhattacharyya coefficient;
- **Correlation functions:** Hamann and Yule's Q coefficients, Pearson's product-moment correlation coefficient, Spearman's and Kendall's rank correlation coefficients, Cosine (if x_i and y_i can have positive and negative values).

3.2 Examples of Complementary (Dis)Similarity Functions

Jaccard (see Sect. 2.1):

$$S_J(x,y) = \frac{a}{a+b+c} = \frac{|X \cap Y|}{|X \cup Y|},$$

$$D_J(x,y) = \frac{b+c}{a+b+c} = \frac{|X \oplus Y|}{|X \cup Y|}.$$

Simple Matching (Sect. 2.1):

$$S_{SM}(x,y) = \frac{a+d}{a+b+c+d} = \frac{|X \cap Y| + |\bar{X} \cap \bar{Y}|}{n},$$

$$D_{SM}(x,y) = \frac{b+c}{a+b+c+d} = \frac{|X \oplus Y|}{n} = \frac{1}{n}\sum\nolimits_1^n |x_i - y_i|.$$

Cosine (Sect. 2.2):

$$cos(x,y) = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2}\sqrt{\sum_{i=1}^n y_i^2}}, x_i, y_i \geq 0, i = 1, \ldots, n.$$

$$D(x,y) = \frac{1}{2}\sum\nolimits_{i=1}^n \left(\frac{x_i}{\sqrt{\sum_{i=1}^n x_i^2}} - \frac{y_i}{\sqrt{\sum_{i=1}^n y_i^2}}\right)^2.$$

Bhattacharyya (Sect. 2.4):

$$S(x,y) = \sum\nolimits_{i=1}^n \sqrt{x_i y_i}.$$

$$D(x,y) = \frac{1}{2}\sum\nolimits_{i=1}^n \left(\sqrt{x_i} - \sqrt{y_i}\right)^2.$$

The last dissimilarity function is called Hellinger discrimination, Matusita measure or Squared-Chord distance.

3.3 Fuzzy Relations and Kleene Algebra of Resemblance Functions

Here, similarity and dissimilarity functions will be also referred to as *resemblance functions*. A resemblance function is a symmetric function $R : \Omega \times \Omega \rightarrow [0,1]$, which is reflexive or irreflexive. We will say that two resemblance functions *have the same type* if both are reflexive or both are irreflexive.

Resemblance function R can be considered as a *fuzzy relation* [2, 17, 28, 29], given by membership function: $\mu_R : \Omega \times \Omega \rightarrow [0,1]$, where $\mu_R(x,y)$ is the strength of the relation R between x and y. The properties of fuzzy relations can be extended on resemblance functions.

Denote $\mathcal{P}(S), \mathcal{P}(D)$ and $\mathcal{P}(R)$ the sets of all similarity, dissimilarity and resemblance functions, respectively, defined on the set Ω. We have $\mathcal{P}(S), \mathcal{P}(D) \subset \mathcal{P}(R)$ and $\mathcal{P}(S) \cup \mathcal{P}(D) = \mathcal{P}(R)$. Define the operations *intersection* $\cap$ and *union* $\cup$ of resemblance functions R_1 and R_2 on the set Ω for all x, y in Ω as follows:

$$(R_1 \cap R_2)(x, y) = min\{R_1(x, y), R_2(x, y)\},$$

$$(R_1 \cup R_2)(x, y) = max\{R_1(x, y), R_2(x, y)\}.$$

The set $\mathcal{P}(R)$ will be a distributive lattice [11], partially ordered by the relation:

$$R_1 \subseteq R_2 \text{ if } R_1(x, y) \leq R_2(x, y) \text{ for all } x, y \text{ in } \Omega.$$

The sets $\mathcal{P}(S)$ and $\mathcal{P}(D)$ are distributive sublattices of $\mathcal{P}(R)$. For all similarity functions $S_1, S_2 \in \mathcal{P}(S)$ and dissimilarity functions $D_1, D_2 \in \mathcal{P}(D)$ it is fulfilled:

$$S_1 \cap S_2, S_1 \cup S_2 \in \mathcal{P}(S), \quad D_1 \cap D_2, D_1 \cup D_2 \in \mathcal{P}(D),$$

$$S_1 \cap D_1 \in \mathcal{P}(D), \quad S_1 \cup D_1 \in \mathcal{P}(S).$$

Consider similarity and dissimilarity functions defined for all x, y in Ω by:

$$D_0(x, y) = 0, \quad D_1(x, y) = \begin{cases} 0, & if\ x = y \\ 1, & otherwise \end{cases}.$$

$$S_0(x, y) = \begin{cases} 1, & if\ x = y \\ 0, & otherwise \end{cases}, \quad S_1(x, y) = 1,$$

D_0 is the least element of $\mathcal{P}(D)$ and $\mathcal{P}(R)$, S_1 is the greatest element of $\mathcal{P}(S)$ and $\mathcal{P}(R)$, S_0 is the least element of $\mathcal{P}(S)$ and D_1 is the greatest elements of $\mathcal{P}(D)$.

The complement $N(R)$ of a resemblance function R is defined for all x, y in Ω by: $N(R)(x, y) = 1 - R(x, y)$. This operation is *involutive*, i.e. for all R in $\mathcal{P}(R)$ we have: $N(N(R)) = R$. The complement of the similarity and dissimilarity functions will be equal to their complementary dissimilarity and similarity functions (1), respectively:

$$N(D) = S, \quad N(S) = D.$$

The lattice $\mathcal{P}(R)$ with the complement N will be a normal De Morgan (Kleene) algebra [5] where for any resemblance functions R_1 and R_2 *De Morgan laws*:

$$N(R_1 \cap R_2) = N(R_1) \cup N(R_2), \quad N(R_1 \cup R_2) = N(R_1) \cap N(R_2),$$

and *normality*:

$$R_1 \cap N(R_1) \subseteq R_2 \cup N(R_2),$$

are fulfilled.

Entropy of Resemblance Functions. On the Kleene algebra of resemblance functions one can introduce a *measure of non-probabilistic entropy* of these functions [5, 9, 15]. Similarity and dissimilarity functions such that $S(x, y) = D(x, y) = 0.5$ for all $x \neq y$ in Ω, have the maximal entropy and the uncertainty of making decision "x and y are similar" is maximal for such functions [9].

3.4 Min-Transitivity and Hierarchical Clustering

A symmetric and reflexive fuzzy relation $S : \Omega \times \Omega \rightarrow [0, 1]$ is called a *fuzzy similarity (fuzzy equivalence) relation* [29] if for all *x, y, z* in Ω it satisfies *min-transitivity*:

$$S(x, z) \geq min\{S(x, y), S(y, z)\}.$$

A dissimilarity function D complementary to min-transitive similarity function is called an *ultrametric* and satisfies for all *x, y, z* in Ω the *ultrametric inequality*:

$$D(x, z) \leq max\{D(x, y), D(y, z)\}.$$

For any $\alpha \in [0, 1]$ the α-cut of fuzzy relation S defines a crisp relation $S_\alpha \subseteq \Omega \times \Omega$ as follows: $S_\alpha = \{(x, y) \in \Omega \times \Omega | S(x, y) \geq \alpha\}$. α-cuts are nested such that from $\alpha > \beta$ it follows $S_\alpha \subseteq S_\beta$.

Optimal and Invariant Hierarchical Clustering. All α-cuts of the min-transitive similarity function (fuzzy equivalence relation) $E : \Omega \times \Omega \rightarrow [0, 1]$ are non-fuzzy equivalence relations; hence they define nested partitions of the set Ω on equivalence classes of these relations. These properties of fuzzy equivalence relations give rise to consider *hierarchical clustering* [21] of the set Ω with similarity function S as a *min-transitive transformation* of this similarity function S into a fuzzy equivalence relation E. Tamura et al. [26] proposed to transform S into its transitive closure $\hat{S}$ that will be fuzzy equivalence relation. It was shown [16, 19], that this method coincides with a spanning tree clustering and with a version of the single linkage hierarchical clustering algorithm.

Batyrshin [4, 6] showed that the solution of the problem of *optimal approximation* of similarity function by fuzzy equivalence relation has the form $E = \widehat{F(S)}$, i.e. it can be presented as the min-transitive closure of similarity function $F(S)$, where F is a "correction" of S such that $F(S) \subseteq S$. In addition, it was studied the problem of construction of *invariant hierarchical clustering algorithms* which are *invariant under monotone transformations of similarity values* and *under initial numbering (indexing) of objects*. The solution of this problem also have been presented in the form $E = \widehat{F(S)}$, where $F(S) \subseteq S$. The parametric family of invariant corrections F has been proposed [4, 6].

3.5 Equivalent Resemblance Functions

Two resemblance functions R_1 and R_2 of the same type defined on the set Ω called *equivalent (by ordering)* [3, 24] if for all *x, y, u, v* in Ω it is fulfilled:

$$R_1(x,y) \le R_1(u,v) \quad \text{if and only if} \quad R_2(x,y) \le R_2(u,v).$$

It is clear that two equivalent resemblance functions should have the same type.

A continuous, strictly increasing function $\varphi : [0,1] \to [0,1]$ such that $\varphi(0) = 0$ and $\varphi(1) = 1$ is called an *automorphism* of the interval $[0,1]$.

Proposition 1. If R is a resemblance function on Ω and φ is an automorphism of the interval [0,1] then the function R_1 defined for all x, y in Ω by:

$$R_1(x,y) = \varphi(R(x,y)),$$

will be a resemblance function equivalent to R.

Below there are examples of simplest equivalent transformations of resemblance functions:

$$R_1(x,y) = R^2(x,y), \quad R_1(x,y) = \sqrt{R(x,y)}.$$

For example, instead of dissimilarity function

$$D(x,y) = \frac{1}{2}\sqrt{\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} - \frac{y_i - \bar{y}}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}\right)^2},$$

one can use the equivalent dissimilarity function

$$D(x,y) = \frac{1}{4}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} - \frac{y_i - \bar{y}}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}\right)^2.$$

Equivalent Resemblance Functions and Invariant Clustering. Equivalence of (dis) similarity functions supposes that the use of such functions in some classification algorithm will give equivalent results. As such clustering algorithms one can use hierarchical clustering algorithms invariant under monotone transformations of similarity values discussed in previous section.

4 Correlation Functions

4.1 Correlation Functions and Correlation Triplets

In Sect. 3 we introduced the definition of correlation function as follows.

A function $A : \Omega \times \Omega \to [-1,1]$ is a *correlation function* (association measure) on Ω if for all x, y in Ω it is *symmetric*:

$$A(x, y) = A(y, x),$$

reflexive:

$$A(x, x) = 1,$$

and negative: $A(x, y) < 0$, for some x, y in Ω.

Correlation function will be called a *weak* correlation function if it does not satisfy the inverse relationship property considered below. Here we consider and extend some results introduced in [10].

Proposition 2. Suppose S and D are similarity and dissimilarity functions on Ω such that for some x, y in Ω it is fulfilled: $S(x, y) < D(x, y)$, then the function defined for all x, y in Ω by:

$$A(x, y) = S(x, y) - D(x, y), \tag{2}$$

is a correlation function. If S and D are complementary then the function A will be a correlation function if for some x, y in Ω it is fulfilled: $S(x, y) < 0.5$.

The obtained formula for A has the reasonable interpretation: *the correlation between x and y is positive if the similarity between them is greater than the dissimilarity, and the correlation is negative in opposite case.*

If the similarity S and dissimilarity D functions are complementary then the correlation function A defined by (2) is called *complementary* to S and D. Complementary functions S, D and A will be denoted as (S, D, A) and called a *correlation triplet*. From the definition of the complementary (dis)similarity functions and from (2) it follows that the similarity, dissimilarity and correlation functions from the correlation triplet (S, D, A) can be obtained one from another for all x, y in Ω as follows:

$$S(x, y) = 1 - D(x, y), \quad D(x, y) = 1 - S(x, y), \tag{3}$$

$$A(x, y) = 2S(x, y) - 1, \quad S(x, y) = \tfrac{1}{2}(A(x, y) + 1). \tag{4}$$

$$A(x, y) = 1 - 2D(x, y), \quad D(x, y) = \tfrac{1}{2}(1 - A(x, y)). \tag{5}$$

4.2 Examples of Constructing Correlation Functions from (Dis)Similarity Functions

Hamann coefficient A_H (see Sect. 2.1). The Simple Matching similarity measure $S_{SM}(x, y) = \frac{a+d}{a+b+c+d}$ has the complementary dissimilarity function $D_{SM}(x, y) = \frac{b+c}{a+b+c+d}$. From (2) we obtain:

$$A(x, y) = S(x, y) - D(x, y) = \frac{a+d}{a+b+c+d} - \frac{b+c}{a+b+c+d} = \frac{(a+d)-(b+c)}{a+b+c+d} = A_H.$$

Yule's Q association coefficient A_{Y-Q} (Sect. 2.1). The function $S_Y(x,y) = \frac{ad}{ad+bc}$ is the similarity function and the function $D_Y(x,y) = \frac{bc}{ad+bc}$ is it's complementary dissimilarity function. From (2) we obtain:

$$A(x,y) = S(x,y) - D(x,y) = \frac{ad}{ad+bc} - \frac{bc}{ad+bc} = \frac{ad-bc}{ad+bc} = A_{Y-Q}.$$

Note that similarly to Yule's Q association and Hamann coefficients it is easy to construct the most of correlation functions considered for binary data [13].

Pearson's product-moment correlation coefficient r (Sect. 2.2). The function

$$D(x,y) = \frac{1}{4}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} - \frac{y_i-\bar{y}}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}\right)^2,$$

is the dissimilarity function. From (5) obtain Pearson's product-moment correlation coefficient:

$$A(x,y) = 1 - 2D(x,y) = 1 - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} - \frac{y_i-\bar{y}}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}\right)^2$$
$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = r.$$

Spearman's rank correlation coefficient ρ (see Sect. 2.3). Consider the function:

$$D(x,y) = \frac{3\sum_{i=1}^{n}(x_i-y_i)^2}{n(n^2-1)}.$$

It satisfies the properties of dissimilarity functions and from (5) we obtain the Spearman's rank correlation coefficient: $A(x,y) = 1 - 2D(x,y) = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = \rho$.

Kendall's rank correlation coefficient τ (Sect. 2.3). Consider the functions:

$$S(x,y) = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} s_{ij}}{n(n-1)/2} = \frac{NC}{n(n-1)/2},$$
$$D(x,y) = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_{ij}}{n(n-1)/2} = \frac{ND}{n(n-1)/2}.$$

They are the complementary similarity and dissimilarity functions, respectively, such that $S(x,y) + D(x,y) = 1$, and from (2) we obtain $A(x,y) = \frac{NC-ND}{n(n-1)/2} = \tau$.

4.3 Strong (Invertible) Correlation Functions on the Sets with Involution Operation

Initially the correlation function (association measure) was defined on the set with involution operation [7] as function satisfying inverse relationship property considered below. Such correlation functions will be called here *strong* or *invertible correlation functions*. It is surprising that all correlation functions considered above are invertible. For this reason, the correlation function which is not invertible will be called a *weak correlation function*.

A function $N : \Omega \to \Omega$ is called a *reflection* or a *negation* on Ω if it satisfies for all x in Ω the *involutivity* property:

$$N(N(x)) = x,$$

and if it is not an identity function, i.e. for some x in Ω it is fulfilled: $N(x) \neq x$.

An element x in Ω such that $N(x) = x$ is called a *fixed point* and the set of all fixed points of the reflection N on Ω is denoted as $FP(N, \Omega)$ or $FP(\Omega)$.

Definition 1. [7] Let N be a reflection on Ω and V be a subset of $\Omega \backslash FP(\Omega)$ closed under N. A *strong correlation function* (association measure) on V is a function $A : V \times V \to [-1, 1]$ satisfying for all x, y in V the properties:

$$\begin{aligned} &A(x, y) = A(y, x), && \text{(symmetry)} \\ &A(x, x) = 1, && \text{(reflexivity)} \\ &A(x, N(y)) = -A(x, y). && \text{(inverse relationship)} \end{aligned}$$

The strong correlation function also will be referred to as an *invertible correlation function*.

Theorem 1. The correlation function A from a correlation triplet (S, D, A) is invertible if and only if the complementary similarity and dissimilarity functions satisfy the following properties:

$$S(x, y) + S(x, N(y)) = 1, \quad D(x, y) + D(x, N(y)) = 1.$$

These properties will be called *bipolarity* properties and corresponding functions S and D will be called *bipolar*, see [8, 10]. The value 1 equals to the sum of the pole values 0 and 1 of the interval [0, 1] of similarity and dissimilarity values. It is clear that S is bipolar if and only if its complementary dissimilarity function D is bipolar.

Similarly, the property of inverse relationship of correlation function can be written in the form of *bipolarity*:

$$A(x, y) + A(x, N(y)) = 0,$$

taking into account that $0 = -1 + 1$, i.e. zero equals to the sum of the pole values of the interval $[-1, 1]$ of correlation values. With this terminology the Theorem 1 can be formulated as follows: *The correlation function A from a correlation triplet (S, D, A) is bipolar if and only if the similarity and dissimilarity functions S and D are bipolar.*

The proof follows, for example, from (4): $A(x,y)+A(x,N(y)) = 2S(x,y)-1+2S(x,N(y))-1 = 2(S(x,y)+S(x,N(y))-1)$, and the first sum equals to zero if and only if the last sum equals to zero.

For complementary (dis)similarity functions we have: $S(x,y)+D(x,y)=1$ and bipolarity of these functions is equivalent to the properties:

$$D(x,y) = S(x,N(y)), \quad S(x,y) = D(x,N(y)). \tag{6}$$

Hence to prove the inverse relationship property of the correlation function A it is sufficient to show the fulfillment of the bipolarity or (6) properties for the (dis)similarity functions S or D used in construction of A by means of (2), (4) or (5).

One can show that all correlation functions considered in Sect. 4.2 are invertible with respect to suitable involutions. See some results in [10].

4.4 Constructing Strong Correlation Functions from Co-symmetric (Dis) Similarity Functions

Similarity S and dissimilarity D functions are *consistent* on the set Ω with involution N if for all x in Ω it is, respectively, fulfilled [9]:

$$S(x,N(x)) = 0, \quad D(x,N(x)) = 1.$$

Resemblance function R is *co-symmetric* on the set Ω with involution N if for all x, y in Ω it is fulfilled [9]:

$$R(N(x),N(y)) = R(x,y).$$

It was shown [7, 9] that a resemblance function R is co-symmetric if and only if for all x, y in Ω it is fulfilled the following property:

$$R(x,N(y)) = R(N(x),y).$$

Proposition 3 [7]. Invertible correlation function is co-symmetric.

Proposition 4 [10]. Bipolar resemblance function is consistent and co-symmetric.

Theorem 1 says that for construction of invertible correlation function from its complementary (dis)similarity functions by (2), (4) or (5) we need to have bipolar (dis) similarity functions. To construct such functions for specific domain is not always easy. From Proposition 4, one can conclude that consistent and co-symmetric (dis)similarity functions may be not so restrictive than bipolar functions, and it is easier to construct such (dis)similarity functions than bipolar functions. The following theorem shows how to use them for constructing invertible correlation functions.

Theorem 2 [7]. Let N be a reflection on Ω and V be a nonempty subset of $\Omega \backslash FP(\Omega)$ closed under reflection N. Let $S : V \times V \to [0,1]$ be a co-symmetric and consistent similarity function, then the function $A : V \times V \to [-1,1]$ defined for all x, y in V by:

$$A(x, y) = S(x, y) - S(x, N(y)), \tag{7}$$

is a strong correlation function on V.

The formula (7) has the simple interpretation: *the correlation between x and y is positive if x is more similar to y than to its negation and the correlation is negative in the opposite case.*

More general methods of constructing invertible correlation functions (association measures) have been proposed in [7, 9, 10]. These methods instead of difference operation in (7) use pseudo-difference operations and instead of consistent similarity functions in (7) they can use similarity functions satisfying weaker conditions.

5 Conclusion and Future Directions of Research

This work presents a short and updated version of a part of author's Lecture on RAAI Summer School 2019. The presented paper can be considered as complementary to the [9]. It includes new methods of construction of correlation functions presented recently on INES 2019 [10] and contains more examples of (dis)similarity and correlation functions illustrating these methods. As a future work, it is supposed to extend the developed approach on other types of association and relationships measures and on other domains.

Acknowledgements. This works partially supported by the project SIP 20196374 IPN and by Organizing Committee of RAAI Summer School. The author thanks all organizers of RAAI Summer School and editors of this book. Special thanks to doctors Gennady Osipov, Alexander Panov and Maria Koroleva.

References

1. Aherne, F.J., Thacker, N.A., Rockett, P.I.: The Bhattacharyya metric as an absolute similarity measure for frequency coded data. Kybernetika **34**, 363–368 (1998)
2. Averkin, A.N., Batyrshin, I.Z., Blishun, A.F., Silov, V.B., Tarasov, V.B.: Fuzzy sets in models of control and artificial intelligence. Pospelov, D.A. (ed.) Nauka, Moscow (1986). (in Russian)
3. Batagelj, V., Bren, M.: Comparing resemblance measures. J. Classif. **12**, 73–90 (1995)
4. Batyrshin, I.Z.: Methods of system analysis based on weighted relations, Ph.D. dissertation. Moscow Power Engineering Institute, Moscow (1982). (in Russian)
5. Batyrshin, I.Z.: On fuzzinesstic measures of entropy on Kleene algebras. Fuzzy Sets Syst. **34**, 47–60 (1990)
6. Batyrshin, I., Rudas, T.: Invariant hierarchical clustering schemes. In: Batyrshin, I., Kacprzyk, J., Sheremetov, L., Zadeh, L.A. (eds.) Perception-Based Data Mining and Decision Making in Economics and Finance, pp. 181–206. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-36247-0_7
7. Batyrshin, I.Z.: On definition and construction of association measures. J. Intell. Fuzzy Syst. **29**, 2319–2326 (2015)

8. Batyrshin, I., Monroy-Tenorio, F., Gelbukh, A., Villa-Vargas, L.A., Solovyev, V., Kubysheva, N.: Bipolar rating scales: a survey and novel correlation measures based on non-linear bipolar scoring functions. Acta Polytechnica Hungarica **14**, 33–57 (2017)
9. Batyrshin, I.: Towards a general theory of similarity and association measures: similarity, dissimilarity and correlation functions. J. Intell. Fuzzy Syst. **36**(4), 2977–3004 (2019)
10. Batyrshin, I.Z.: Constructing correlation coefficients from similarity and dissimilarity functions. In: INES 2019, IEEE 23rd IEEE International Conference on Intelligent Engineering Systems, Hungary, 25–27 April. IEEE, Gödöllő (2019)
11. Birkhoff, G.: Lattice Theory, 3rd edn. American Mathematical Society, Providence (1967)
12. Chen, P.Y., Popovich, P.M.: Correlation: Parametric and Nonparametric Measures. Sage, Thousand Oaks (2002)
13. Choi, S.S., Cha, S.H., Charles, C.T.: A survey of binary similarity and distance measures. J. Syst. Cybern. Inform. **8**, 43–48 (2010)
14. Clifford, H.T., Stephenson, W.: An Introduction to Numerical Classification. Academic Press, New York (1975)
15. De Luca, A., Termini, S.: A definition of a nonprobabilistic entropy in the setting of fuzzy sets. Inform. Control **20**, 301–312 (1972)
16. Dunn, J.C.: A graph theoretic analysis of pattern classification via Tamura's fuzzy relation. IEEE Trans. Syst. Man Cybern. **3**, 310–313 (1974)
17. Fodor, J.C., Roubens, M.R.: Fuzzy Preference Modelling and Multicriteria Decision Support, vol. 14. Springer, Dordrecht (1994). https://doi.org/10.1007/978-94-017-1648-2
18. Gibbons, J.D., Chakraborti, S.: Nonparametric Statistical Inference, 4th edn. Dekker, New York (2003)
19. Gower, J.C., Ross, G.J.S.: Minimum spanning trees and single linkage cluster analysis. Appl. Stat. **18**, 54–64 (1969)
20. Janson, S., Vegelius, J.: Measures of ecological association. Oecologia **49**, 371–376 (1981)
21. Johnson, S.C.: Hierarchical clustering schemes. Psychometrika **32**, 241–254 (1967)
22. Kendall, M.G.: Rank Correlation Methods, 4th edn. Griffin, London (1970)
23. Legendre, P., Legendre, L.F.: Numerical Ecology, 2nd edn. Elsevier, Amsterdam (1998). English edn.
24. Lesot, M-J., Rifqi, M., Benhadda, H.: Similarity measures for binary and numerical data: a survey. Int. J. Knowl. Eng. Soft Data Paradigms **1**, 63–84 (2009)
25. Rauschenbach, G.V.: Proximity and similarity measures. In: Analysis of Non-Numerical Information in Sociological Research, Nauka, Moscow, pp. 169–202 (1985). (in Russian)
26. Tamura, S., Higuchi, S., Tanaka, K.: Pattern classification based on fuzzy relations. IEEE Trans. Syst. Man Cybern. **1**, 61–66 (1971)
27. Tan, P.N., Kumar, V., Srivastava, J.: Selecting the right interestingness measure for association patterns. In: 8th Proceedings of Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 32–41 (2002)
28. Zadeh, L.A.: Fuzzy sets. Inf. Control **8**, 338–353 (1965)
29. Zadeh, L.A.: Similarity relations and fuzzy orderings. Inf. Sci. **3**, 177–200 (1971)