Correlation Analysis of Binary Similarity and Dissimilarity Measures

by
Seung-Seok (Seung) Choi

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Professional Studies
in Computing

at

Seidenberg School of Computer Science and Information Systems

Pace University

August 18, 2008

We hereby certify that this dissertation, submitted by Seung-Seok Choi, satisfies the dissertation requirements for the degree of *Doctor of Professional Studies in Computing* and has been approved.

_____-_____

Dr. Sung-Hyuk Cha Date
Chairperson of Dissertation Committee

_____-_____

Dr. Charles Tappert Date
Dissertation Committee Member

_____-_____

Dr. Ronald Frank Date
Dissertation Committee Member

Seidenberg School of Computer Science and Information Systems
Pace University 2008

Abstract

Correlation Analysis of Binary Similarity and Dissimilarity Measures

by
Seung-Seok (Seung) Choi

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Professional Studies
in Computing

August 18, 2008

There are numerous binary similarity measures. Different binary similarity measures estimate different aspects of taxonomic relationships between two objects. This study presents the correlation of 76 binary similarity and dissimilarity measures used in many different fields. These measures are then grouped on the basis of their synthetic properties, arithmetic relationships, and chronological order. Five binary data sets from three different binary types are used to investigate the data dependency and the data invariance: a hypothetical random binary data set, three different sets of hypothetical equal random binary data set (the first set having 50% of 1s, the second 90% of 1s, and the third 90% of 0s), and a flattened nominal mushroom data set. This is the most extensive study of binary similarity measures ever conducted, analyzing the correlations of 2,850 pairs of measures, and comparing them in the five different data sets. Using the hierarchical clustering technique with the agglomerative single linkage, the 76 binary similarity and dissimilarity measures are clustered by their similarity values. In addition, the correlations of the binary similarity measures are quantified based on the similarity values computed from each pair. A variety of correlation patterns are discovered, and they are presented as 12 different correlation patterns as the data set dependent correlation and 3 types of correlation patterns as the data set invariant correlation. The correlation matrix representing 2,850 pairs of correlations is transformed to a gray-scaled square image for improved visualization and better understanding. Finally, the statistical significance tests are performed for the correlation matrices obtained from the five data sets. The distribution curves on each data set show that the correlations of binary similarity and dissimilarity measures are affected by the data set domains.

Keywords: binary similarity measure, distance measure, correlation, hierarchical cluster analysis

Acknowledgements

I would like to acknowledge and extend my heartfelt gratitude to the following persons who have made the completion of this dissertation possible;

Drs. Sung-Hyuk Cha, Charles Tappert, Fred Grossman, and Ron Frank for the constant encouragement, the vital inspiration, and the helpful comments.

I would like to thank most especially to my family; Sinyeon Kee, my wife, Justin K. Choi, my son, Jung-Hee Kang, my mother, Bosoon Kim, my mother-in-law, and Hee-Kwan Kee, my father-in-law, who had fully supported me through this dissertation.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

From the scientific, or mathematical, point of view, *similarity* can be defined as a quantitative degree of how similar two objects are. The synonyms for *similarity* include *resemblance, affinity, association*, and *proximity* [76]. Therefore, a *similarity measure* estimates the degree of similarity or resemblance between two objects, thus, it can be defined as a quantitative estimation of the similarity between two objects. The *similarity measures* are often called the *similarity coefficients*. Unlike the similarity measures, the *dissimilarity measures* provide distance values. *Distance* is a quantitative degree of how different two objects are. Therefore, the dissimilarity measures estimate the degree of distance or difference between two objects. The *dissimilarity measures* are often called the *distance measures* or *diversity indices*. Since the distance, or dissimilarity, values are the complement of similarity values, many similarity coefficients can identically be transformed to dissimilarity coefficients. The conventional definitions of the *similarity measures,* thus, often include the *dissimilarity measures*, such as Euclidean distance and Hamming distance [39]. The *binary similarity measures* and *binary dissimilarity measures* are used especially for binary data in which the attributes are either 1 or 0. Each binary similarity measure has its own properties and features. Different binary similarity measures estimate different aspects of taxonomic relationships between two objects. Some measures only account for positive matches while some include negative matches,

some apply weights on either matches or mismatches, or both. Some are monotonically related to others while some are not related to others at all.

The representation of objects differs depending on its variable type. That is, a variable is any measured characteristic or attribute that differs for different objects. The most widely used data representations include *nominal variables*, *ordinal variables*, *interval variables*, and *binary variables* [93, 94]. Binary variables may be *symmetric* or *non-symmetric* (*asymmetric).*

The binary similarity and dissimilarity measures play a critical role in classification problems, cluster analysis, and identification issues because the use of an appropriate measure produces a more accurate analysis. Over a hundred years, many researches have taken elaborate efforts to find the most meaningful binary similarity measures in their fields. For example, the Jaccard similarity measure was used for clustering ecological species [43] which was the first implementation of the binary similarity measure, and Forbes proposed a similarity coefficient for clustering ecologically related species [28, 29]. After that, they were subsequently applied in biology [42, 56], ethnology [20], taxonomy [76], image retrieval [73], geology [57], and chemistry [84]. In addition, they have been actively used to solve the identification problems in biometrics such as fingerprint [85], iris images [11], and handwriting character images [9, 10]. Numerous similarity measures and dissimilarity measures for binary data have been proposed in those areas. [16, 32, 33, 34, 35, 42, 49, 74] discuss their properties and features. Furthermore, many of them have been developed and implemented for real applications.

For example, SPSS CLUSTER (SPSS 2001) provides 27 binary similarity or dissimilarity measures implemented [90].

The inclusion or exclusion of *negative matches* in the binary similarity measures have been a continuous issue [22, 27, 30, 32, 33, 34, 35, 74, 76]. The meaning of *positive matches* and *negative matches* is from the combination of the two states (presence or absence) of features between two binary data. A *negative match* is where the two binary instances have 0 (or absence) as their value while a *positive match* is where the two binary instances have 1 (or presence) as their value.

When features are absent in both objects, should we consider it as important as when they are present in both data? The following 'wing' example explains the difficulties over the appropriate way to deal with negative matches when estimating the similarity value.

> *"The absence of wing, when observed among a group of distantly related organisms (such as camel, louse and nematode), would surely be an absurd indication of affinity. Yet a positive character such as the presence of wings (or flying organs defined without qualifications as to kind of wing) could mislead equally when considered for a similarity heterogeneous assemblage (for example, bat, heron and dragonfly). Neither can we argue that absence of a character may due to a multitude of causes and that matched absence in a pair of OTUs is therefore not 'true resemblance', for, after all, we know little more about the origins of matched positive characters"* [76]

This problem has been argued in [22, 30, 74, 76]. Sokal et al. argued that the negative matches do not mean necessarily any similarity between two objects in [76]. This is because an almost infinite number of attributes is possibly lacking in two objects. In cases where the two binary states are not equally important, such as in the asymmetric type of binary data, the positive matches are usually more significant than the negative matches. However, they prefer to inclusion of negative matches, pointing out that most studies in biology since 1957 had used the similarity coefficients including negative matches. Moreover, Sokal et al. insisted on exclusion of some positive matches when they are invariant in the data set. They pointed out that it is not adequate to use positive attributes that are invariant over the entire sample, that is, the attributes that are possessed by all samples [76].

The binary similarity measures such as the Dice & Sorenson measure or the Pearson measure include both negative and positive matches in computing their similarity coefficient, while others do not consider negative matches in their similarity measurement. It is interesting to observe that all dissimilarity measures, such as the Hamming distance or the Lance & Williams coefficient, intrinsically include negative matches for computing their distance values.

Even though the numerous binary similarity measures have been described in the literatures, only a few comparative studies collected the wide variety of binary similarity measures [11, 42, 44, 81, 85, 89]. In [42], Hubalek collected 43 similarity measures, and 20 of them were used for cluster analysis on fungi data to produce five clusters of related

coefficients. He demonstrated that coefficients from different clusters yielded different dendrograms and grouped them into two types of similarity measures, the correlation based similarity measures and the non-probabilistic similarity measures. The former including *the Pearson I, the Pearson II, the Pearson III, the Cole, the Pearson & Heron I, the Yule Q, the Yule w, the Michael, the Forbes II, the Tarwid, the Dennis, the Stiles,* are limited in measuring similarity but highly applicable for assessing an association. The latter such as the *Jaccard, the Dice & Sorenson, the Kulczynski I, and the Sokal & Michener* are good for measuring both similarity and association [42]. In [44], ecological data on 25 fish species were analyzed with eight similarity coefficients: *the Jaccard, the Ochiai, the Phi, the Rogers-Tanimoto, the Russell and Rao, the Simple Matching, the Sorensen-Dice, and the Yule* similarity coefficients. These coefficients were divided into two groups, one for representing the co-occurrence and the other for measuring their association. *The coefficients of co-occurrence* (i.e., *the Jaccard, the Rogers & Tanimoto, the Russell & Rao, the Simple Matching, and the Dice & Sorensen*) incorporate information associated with the frequency of occurrence while *the measures of association* incorporate implicit centering transformations that reduce the size influence associated with the frequency of occurrence. Jackson et al. observed that the measures of association (i.e., *the Phi and the Yule*) were less affected to the frequency of occurrence [44]. Tubbs summarized seven conventional similarity measures to solve the template matching problem [81], and Zhang et al. compared those seven measures to show the recognition capability in handwriting identification [89]. Willett evaluated 13 similarity measures for binary fingerprint code [85]. Cha et al. proposed weighted binary measurement to improve classification performance based on the comparative study [11].

This study has five key contributions. First, it is the most extensive comparative study of binary similarity measures, collecting 76 binary similarity and dissimilarity measures developed over a hundred years, and analyzing the correlations of 2,850 pairs of those measures. Second, the measures are grouped on the basis of their synthetic properties, arithmetic relationships, and chronological order. Third, a variety of correlation patterns are discovered and summarized into 12 different correlation patterns for the data set dependent correlation of binary similarity measures and 3 types of correlation patterns for the data set invariant correlation of binary similarity measures. Fourth, the correlation matrix representing 2,850 pairs of correlations among the 76 binary similarity and dissimilarity measures is transformed to a gray-scaled square image for improved visualization and understanding. Fifth, the distribution curves obtained from a statistical significance test on five binary data sets show that the binary similarity and dissimilarity measures behave differently depending on the data set.

Operational Taxonomic Units (OTUs) [22] are used to express the definitions of the 76 binary similarity and dissimilarity measures collected, summarizing the similarity values between two binary objects in a 2 x 2 contingency table. All binary measures are reviewed and evaluated by their conceptual relationships.

A synthetic grouping was performed on the basis of their common properties in order to present the relationships among them. Two aspects are considered for the grouping. One aspect is inclusion or exclusion of negative matches, that is, the binary similarity or

dissimilarity coefficients can be grouped on the basis of how they deal with the negative matches. The second aspect is how the similarity values are computed. Some coefficients are based on the frequencies of occurrence, while others measure the correlation between two binary objects. They are categorized into the following three groups: *Featured based Binary Similarity Measures*, *Correlation based Binary Similarity Measures*, and *Binary Dissimilarity Measures*. The feature-based binary similarity measures and the correlation-based binary similarity measures are subcategorized into *negative matches inclusive* and *negative matches exclusive*.

The feature-based binary similarity measures are designed to express the proportion of matching and mismatching in common between two binary instances. 21 binary similarity measures such as the *Jaccard,* the *Tanimoto,* the *Sokal & Michener,* the *Russell & Rao,* the *Intersection,* the *Dice & Sorenson,* the *Czekanowski,* the *Sokal & Sneath I, II, III,* the *Roger & Tanimoto,* the *Kulczynski I,* the *Hamann,* the *Dispersion,* the *Baroni-Urbani & Buser I, II,* the *Faith,* the *Nei & Lei,* the *Gower & Legendre* are included in this category.

The binary dissimilarity measures estimate distance instead of similarity between two binary objects. This group includes 14 binary measures such as the *Hamming,* the *Binary Euclidean,* the *Binary Squared Euclidean,* the *Manhattan,* the *Mean Manhattan,* the *City Block,* the *Size Difference,* the *Pattern Difference,* the *Shape Difference,* the *Variance,* the *Minkowski,* the *Canberra, Lance & Williams, and* the *Bray-Curtis*.

The correlation based binary similarity measures estimate the likelihood that an observed number of shared attributes between two binary instances arose by chance, which is the degree that two instances are associated. 41 binary similarity measures are grouped in this

category: the *Sokal & Sneath IV, V,* the *Kulczynski II,* the *Cosine,* the *Ochiai I, II,* the *Goodman & Kruskal,* the *Anderberg,* the *Yule Q,* the *Yule w,* the *Pearson & Heron I (Phi),* the *Pearson & Heron II,* the *Pearson I (Chi-square),* the *Pearson II, III,* the *Gower,* the *Forbes I, II,* the *Fossum,* the *Simpson,* the *Stiles,* the *Dennis,* the *McConnaughey,* the *Braun-Banquet,* the *Sorgenfrei,* the *Fager & McGowan,* the *Mountford,* the *Peirce,* the *Eyraud,* the *Michael,* the *Gilbert & Wells,* the *Tarwid,* the *Otsuka,* the *Hellinger,* the *Chord,* the *Tarantula,* the *AMPLE,* the *Driver & Kroeber,* the *Johnson, and* the *Cole*.

In addition to the synthetic grouping, the arithmetic grouping, in which the measures are categorized by their denominators and numerators, and the chronological grouping are addressed. An historical overview helps us to understand their evolution and expansion. For example, the binary similarity coefficients proposed by Peirce, Yule, and Pearson in 1900s contributes to the evolution of the various correlation based binary similarity measures. The Euclidean distance measure becomes a milestone of the dissimilarity measures such as the Hamming distance or the Lance & Williams distance measures. The Jaccard coefficient proposed in 1901, which becomes a standard of the negative matches exclusive binary similarity measure, is still widely used in various fields such as ecology and biology.

The correlations among the 76 binary measures are analyzed using all of the 2,850 possible pairs. The correlation coefficient is used to quantify the relationships of each pair of similarity values computed from the binary similarity measures. Using the hierarchical clustering technique with the agglomerative single linkage with the average

clustering method [18], 76 binary similarity and dissimilarity measures are clustered by their similarity values. Five different data sets are used in this study: a hypothetical *random binary data set*, three different sets of hypothetical *equal random binary data set* (the first set having 50% of 1s, the second 90% of 1s, and the third 90% of 0s), and the *flattened nominal mushroom data set.*

The observations are focused on two aspects. The first is focused on how the correlations of binary similarity measures diversified between hypothetical random binary data and domain specific real data. The second is focused on how the inclusion or exclusion of negative matches affects the similarity values on the data sets in which the number of two states, 1 and 0, are unequally assigned.

As a result, various shapes of correlation curves are found which include (1) Straight Linear, (2) Reversed Linear, (3) Convex, (4) Concave, (5) Spread, (6) V Shape, (7) U Shape, and (8) L shape. The more semantically similar the two binary similarity measures are, the clearer linear shape they have, that is, the straight linear shape of correlation curve represents the good fitness between two binary similarity measures. The reversed linear shape represents the good fitness between a binary similarity measures and a binary distance measures. The convex and concave shapes are usually found in relationship between two correlation-based binary similarity measures. V, U or L shapes are found in the relationships between a feature-based binary similarity measure and a correlation-based binary similarity measure.

A cluster analysis of the binary similarity and dissimilarity measures is performed by averaging 30 independent trials for each of the five data sets. This clustering resulted in 11 groups in the random binary data set, 7 groups in the equal random binary data set having 50% 1s and 50% 0s, 7 groups in the equal random binary data set having 90% 1s and 10% 0s, 10 groups in the equal random binary data set having 10% 1s and 90% 0s, and 8 groups in the flattened nominal mushroom data set. The correlation patterns among the three different sets of equal random binary data are generally similar while the difference of the correlation patterns between the random binary data set and the flattened nominal mushroom data set significantly different.

We discuss two aspects of correlations changes: the *Data Set Dependent Correlation* and the *Data Set Invariant Correlation*. *Data Set Dependent Correlation* means that the correlations between two binary similarity measures are changed when the different types of data are used. *Data Set Invariant Correlation* means that the correlations between two binary similarity measures are not changed even though different types of data set are applied. Substantial correlations changes are also observed in the some pairs of binary similarity measures, such as a pair of *the Hamming* and *the Anderberg*, a pair of *the Ochiai I* and *the Stiles*, or a pair of *the Jaccard* and *the Tanimoto*, etc. 12 different patterns for the data set dependent correlation are addressed. For the data set invariant correlations, 3 different correlation patterns are found within the same data sets. They include (1) a diagonal straight line, (2) a reversed diagonal straight line, and (3) a convex curve.

The first pattern can be found in any pairs involving *the Johnson, the McConnaughey, the Ochiai I, the Cosine, the Nei & Li, the Sokal & Sneath I, and the Kulczynski II,* any pairs of *the Sokal & Sneath II, the Sokal & Sneath IV, the Sokal & Sneath V, the Ochiai II, the Gower, the Forbes II, the Cole, the Pearson & Heron I, the Pearson III, the Dennis, the Roger & Tanimoto, the Sokal & Michener, and the Hamann,* a pair of *the Pearson I and the Pearson II, and a pair of the Baroni-Urbani & Buser I and the Baroni-Urbani & Buser II.*

The second pattern is from the correlation between a similarity measure and a dissimilarity measure which of complement is identical with the similarity measure. This type of correlation can be found in the pairs of similarity measures such as *the Sokal & Sneath II, the Sokal & Sneath IV, the Sokal & Sneath V, the Ochiai II, the Gower, the Forbes II, the Cole, the Pearson & Heron I, the Pearson III, the Dennis, the Roger & Tanimoto, the Sokal & Michener, and the Hamann* and dissimilarity measures such as *the Binary Euclidean, the Shape Difference, the Dispersion, the Mean Manhattan, the Canberra, the Variance, the Minkowski, and the Hamming* and the pairs of the similarity measures such as *the Johnson, the McConnaughey, the Ochiai I, the Cosine, the Nei & Li, the Sokal & Sneath I, and the Kulczynski II* and the dissimilarity measures such as *the Bray & Curtis, the Lance & Williams, the Hellinger, and the Chord*.

The third pattern can be found in the pair of *the Stiles and the Pearson I or the Pearson II* and the pair of *the AMPLE and the Pearson I or the Pearson II* in which their correlation, r is always established between 0.04 -0.05 ranges in the all data sets.