A Survey of Binary Similarity and Distance Measures

Seung-Seok Choi, Sung-Hyuk Cha, Charles C. Tappert
Department of Computer Science, Pace University
New York, US

ABSTRACT

The binary feature vector is one of the most common representations of patterns and measuring similarity and distance measures play a critical role in many problems such as clustering, classification, etc. Ever since *Jaccard* proposed a similarity measure to classify ecological species in 1901, numerous binary similarity and distance measures have been proposed in various fields. Applying appropriate measures results in more accurate data analysis. Notwithstanding, few comprehensive surveys on binary measures have been conducted. Hence we collected 76 binary similarity and distance measures used over the last century and reveal their correlations through the hierarchical clustering technique.

Keywords: binary similarity measure, binary distance measure, hierarchical clustering, classification, operational taxonomic unit

1. INTRODUCTION

The binary similarity and dissimilarity (distance) measures play a critical role in pattern analysis problems such as classification, clustering, etc. Since the performance relies on the choice of an appropriate measure, many researchers have taken elaborate efforts to find the most meaningful binary similarity and distance measures over a hundred years. Numerous binary similarity measures and distance measures have been proposed in various fields.

For example, the *Jaccard* similarity measure was used for clustering ecological species [20], and *Forbes* proposed a coefficient for clustering ecologically related species [13, 14]. The binary similarity measures were subsequently applied in biology [19, 23], ethnology [8], taxonomy [27], image retrieval [25], geology [24], and chemistry [29]. Recently, they have been actively used to solve the identification problems in biometrics such as fingerprint [30], iris images [4], and handwritten character recognition [2, 3]. Many papers [7, 16, 17, 18, 19, 22, 26] discuss their properties and features.

Even though numerous binary similarity measures have been described in the literature, only a few comparative studies collected the wide variety of binary similarity measures [4, 5, 19, 21, 28, 30, 31]. Hubalek collected 43 similarity measures, and 20 of them were used for cluster analysis on fungi data to produce five clusters of related coefficients [19]. Jackson et al. compared eight binary similarity measures to choose the best measure for ecological 25 fish species [21]. Tubbs summarized seven conventional similarity measures to solve the template matching problem [28], and Zhang et al. compared those seven measures to show the recognition capability in handwriting identification [31]. Willett evaluated 13 similarity measures for binary fingerprint code [30]. Cha et al. proposed weighted binary measurement to improve classification performance based on the comparative study [4].

Few studies, however, have enumerated or grouped the existing binary measures. The number of similarity or dissimilarity measures was often limited to those provided from several commercial statistical cluster analysis tools. We collected and analyzed 76 binary similarity and distance measures used over the last century, providing the most extensive survey on these measures.

This paper is organized as follows. Section 2 describes the definitions of 76 binary similarity and dissimilarity measures. Section 3 discusses the grouping of those measures using hierarchical clustering. Section 4 concludes this work.

2. DEFINITIONS

Table 1 OTUs Expression of Binary Instances i and j

j \ i	1 (Presence)	0 (Absence)	Sum
1 (Presence)	$a = i \bullet j$	$b = \bar{i} \bullet j$	$a+b$
0 (Absence)	$c = i \bullet \bar{j}$	$d = \bar{i} \bullet \bar{j}$	$c+d$
Sum	$a+c$	$b+d$	$n=a+b+c+d$

Suppose that two objects or patterns, i and j are represented by the binary feature vector form. Let n be the number of features (attributes) or dimension of the feature vector. Definitions of binary similarity and distance measures are expressed by *Operational Taxonomic Units* (OTUs as shown in Table 1) [9] in a 2 x 2 contingency table where a is the number of features where the values of i and j are both 1 (or presence), meaning '*positive matches*', b is the number of attributes where the value of i and j is (0,1), meaning 'i absence *mismatches*', c is the number of attributes where the value of i and j is (1,0), meaning 'j absence mismatches', and d is the number of attributes where both i and j have 0 (or absence), meaning '*negative matches*'. The diagonal sum $a+d$ represents the total number of matches between

CHAPTER 6

Distance Measures

Background

The first step of most multivariate analyses is to calculate a matrix of distances or similarities among a set of items in a multidimensional space. This is analogous to constructing the triangular "mileage chart" provided with many road maps. But in our case, we need to build a matrix of distances in hyperspace, rather than the two-dimensional map space. Fortunately, it is just as easy to calculate distances in a multidimensional space as it is in a two-dimensional space.

This first step is extremely important. If information is ignored in this step, then it cannot be expressed in the results. Likewise, if noise or outliers are exaggerated by the distance measure, then these unwanted features of our data will have undue influence on the results, perhaps obscuring meaningful patterns.

Distance concepts

Distance measures are flexible:

- Resemblance can be measured either as a distance (dissimilarity) or a similarity.
- Most distance measures can readily be converted into similarities and vice-versa.
- All of the distance measures described below can be applied to either binary (presence-absence) or quantitative data.
- One can calculate distances among either the rows of your data matrix or the columns of your data matrix. With community data this means you can calculate distances among your sample units (SUs) in species space or among your species in sample space.

Figure 6.1 shows two species as points in sample space, corresponding to the tiny data set below (Table 6.1). We can also represent sample units as points in species space, as on the right side of Figure 6.1, using the same data set.

There are many distance measures. A selection of the most commonly used and most effective measures are described below. It is important to know the domain of acceptable data values for each distance measure (Table 6.2). Many distance measures are not compatible with negative numbers. Other distance measures assume that the data are proportions ranging between zero and one, inclusive.

Table 6.1. Example data set. Abundance of two species in two sample units.

	Species	
Sample unit	1	2
A	1	4
B	5	2

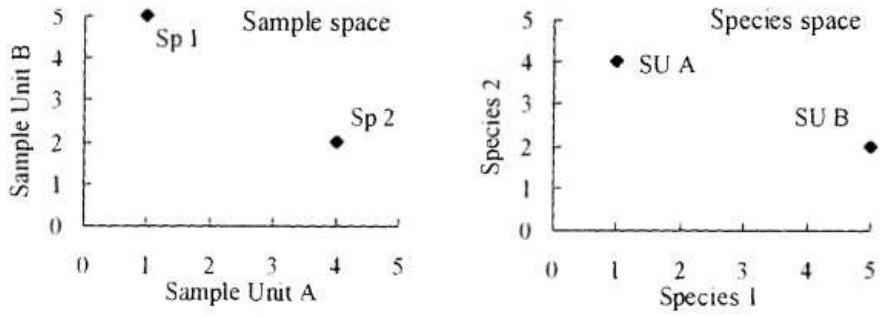

Figure 6.1. Graphical representation of the data set in Table 6.1. The left-hand graph shows species as points in sample space. The right-hand graph shows sample units as