A Standardized Treatment of Binary Similarity Measures with an Introduction to k-Vector Percentage Normalized Similarity

Brian Stacey

1 August 2016

A Standardized Treatment of Binary Similarity Measures with an Introduction to *k*-Vector Percentage Normalized Similarity

Brian Stacey

Abstract

This paper attempts to codify a standard nomenclature for similarity measures based on recent literature and to advance the field of similarity measures through the introduction of non-binary similarity between more than two attribute vectors.

JEL Classifications: C14, C53, C65

Keywords: Binary Similarity, Nonbinary Similarity, Nonparametric Similarity Testing, Multivector Similarity

A Standardized Treatment of Binary Similarity Measures with an Introduction to k-Vector Percentage Normalized Similarity

Numerous papers have been written detailing methods to measure the similarity of two or more vectors (series) of binary attributes. This paper attempts to codify a standard nomenclature for similarity measures based on recent literature and to advance the field of similarity measures through the introduction of non-binary similarity between more than two attribute vectors. The first part of the paper, following the literature review, introduces the standardized nomenclature, while the second part builds the case for similarity measures of percentage normalized attributes, both in two and k-vector formulations. This includes a proposed method for evaluating any size group of vectors. For a detailed discussion concerning the application of the many similarity measures, how they are derived, and the similarities and differences between them Matthijs Warrens' 2008 paper: "Similarity Coefficients for Binary Data" is a great resource. Warrens provides a thorough treatment of the subject through multi-variable measures and where a more detailed understanding of certain measures is desired, this paper will defer to his.

Literature Review

As stated above Warrens (2008) provides a thorough treatment and remains the go-to reference for similarity measures. Warrens details the relationship between the different families of binary similarity metrics and generalizes most to a *k*-vector model. The treatment of non-binary measures is limited to basic distance measures (dissimilarity), and doesn't delve much beyond discussing Euclidean distance as the complement to Sokal-Michener (simple matching) for binary data (2 vector only); indeed, Warrens refers his readers to other sources for non-binary treatments. Warrens does include several interesting proofs. Especially relevant to this paper is

his proof concerning the relative value of various averages; specifically a comparison of the various averages of S_{Dice1} (a/p_1) and S_{Dice2} (a/p_2) shows that the square of the geometric mean (Sorgenfrei) is always less than Jaccard, which is always less than the minimum, harmonic mean, geometric mean, arithmetic mean, and maximum in that order. This holds for all formulations, and helps inform the decision concerning whether to use the arithmetic or quadratic mean when later discussing non-binary *k*-vector percentage normalized metrics. Warren also provides a detailed explanation for methods employed to correct for chance agreement between two vectors. It should be possible to employ these correction methods with *k*-vector percentage normalized non-binary measures, however that assertion is not tested in this paper.

Whereas Warrens is the most detailed, Choi, Cha, & Tappert (2010) is easily the most accessible with regard to explaining the four cases of agreement/disagreement among two vectors. Choi, et al develop a fairly comprehensive list of formulations for common similarity measures and is one of the best first resources for any practitioner or novice in the field.

Warrens and others (Choi, Cha, & Tappert 2010, Lourenço, Lobo, & Bação 2006) use a somewhat standardized terminology for the various cases of agreement or disagreement between two vectors. Warrens switches back and forth between using each term to indicate the raw value (count of cases of agreement) versus the arithmetic average (cases of agreement divided by number of attributes *n*). Although he attempts to indicate each time which he is using, and often the formulation results in their being no difference in the outcome, the few times that it does matter become especially confusing. To address this shortfall this paper uses *a,b,c,d* for raw values and *a′,b′,c′,d′* for the arithmetic average over *n*; this notation applies for all levels of discussion, and each time the prime version of a measure is seen it is the arithmetic average over *n* of the raw measure.

Zhang and Srihari (2003) lay a fair groundwork for treating similarity as the complement to distance (or dissimilarity) and make one of the clearer cases for defining a measure as metric if it meets the four criteria of: non-negativity, commutativity, reflexivity, and satisfying the triangle inequality. Other researchers (e.g. Warrens 2008) focus on the triangle inequality and rarely mention reflexivity.

Several authors approach similarity as either a function of sets or with a Bayesian approach (Novak & Pap 2012, Lourenço, Lobo, & Bação 2006). DeSarbo, De Soete, & Eliashberg (1987) posit that similarity measures can be treated as a Probit regression model based on the result being a probability of agreement between vectors. These methods are not explored here. However Novak & Pap's discussion of a similarity between a single vector and *k*-1 other vectors forms the basis for the *k*-vector approach discussed later.

Lastly, Wilson and Martinez (1997) establish the Heterogeneous Euclidean-Overlap Metric (HEOM) which forms the basis for the decision model posited as part of the k-vector percentage normalized metric. They also do a fair job of explaining the various distance measures with regard to averaging and how they interrelate (especially Minkowsky versus all others) .

There exists a veritable cornucopia of thoughtful analyses concerning measures of distance and similarity, especially when limited to binary attributes among two vectors. The depth of analysis decreases with the introduction of *k*-vector formulations, and the analysis is nearly non-existent for non-binary *k*-vector formulations. Discussion of normalization is almost completely limited to Wilson and Martinez (1997), and even then it is somewhat of a side note.

Binary Measures

Among the many papers written on this subject, most authors, especially of the more recent treatments, have tacitly agreed upon a somewhat standardized nomenclature, however there is some variation, especially when evaluating the defined measures as the arithmetic average over *n*. Warrens (2008) and Choi, et al (2010) interchangeably use *a*, *b*, *c*, and *d* to indicate the raw values of those measures and the averaged versions of the same.

Two Vector Definitions

Suppose there are two objects (X_1 and X_2) each defined by a series of binary attributes such that they could be expressed as:

	X_1	X_2
Attribute 1	1	1
Attribute 2	1	0
Attribute 3	0	1
Attribute 4	0	0
…		
Attribute n	1	1

Table 1: Binary Attributes (values are to show the possible options)

Where 1 indicates the presence of the attribute. There exist within Table 1, four distinct situations, defined as:

	$X_1=1$	$X_1=0$	
$X_2=1$	a	c	$a+c=p_2$
$X_2=0$	b	d	$b+d=q_2$
	$a+b=p_1$	$c+d=q_1$	$a+b+c+d=n$

Table 2: Attribute Measures

Utilizing the measures defined in Table 2 it can be seen from Table 1 that these values can be defined in several ways; for a single attribute, *a* can be defined as: the result of the Boolean

expression $a = X_{1,j} \times X_{2,j}$, the arithmetic expression $a = X_{1,j} * X_{2,j}$, or in set notation as $a = X_{1,j} \cap X_{2,j}$. When expanded to include all attributes it becomes $a = \sum_{j=1}^{n}(X_{1,j} * X_{2,j})$. Including all measures yields:

	Boolean	Arithmetic	Set
a=	$X_1 \times X_2$	$\sum X_1 * X_2$	$X_1 \cap X_2$
b=	$X_1 \times \overline{X_2}$	$\sum X_1 * (1 - X_2)$	$X_1 \backslash X_2$
c=	$\overline{X_1} \times X_2$	$\sum (1 - X_1) * X_2$	$X_2 \backslash X_2$
d=	$\overline{X_1} \times \overline{X_2}$	$\sum (1 - X_1) * (1 - X_2)$	$U \backslash (X_1 \cup X_2)$
p_1=	X_1	$\sum X_1$	X_1
p_2=	X_2	$\sum X_2$	X_2
q_1=	$\overline{X_1}$	$\sum (1 - X_1)$	$U \backslash (X_1)$
q_2=	$\overline{X_2}$	$\sum (1 - X_2)$	$U \backslash (X_2)$
n=	$X_1 + \overline{X_1} = X_2 + \overline{X_2}$	$p_1 + q_1 = p_2 + q_2 = a + b + c + d$	$(X_1 \cup X_2) - (X_1 \cap X_2)$

Table 3: Definitions of Measures

The formulation for *n* in Table 3 assumes that the number of attributes measured for each X is the same (if they weren't the same, the whole concept of a similarity measure falls apart), this value can easily be found within R using length(X_1).

For those values used by Warrens (2008) and Choi, et al (2010) where the measure is averaged over *n*, the nomenclature *a′* (or *b′*, *c′*, *d′*, etc.) will be used going forward, such that:

$a' = \frac{a}{n} = \frac{a}{a+b+c+d}$

This reduces several existing similarity measures to single character expression, e.g. $S_{Russell\&Rao} = a'$. Appendix A contains several common similarity measures in 2 and *k*-vector formulations.

At this point an important heuristic should be presented regarding the choice between measures that account for *d* (co-non-occurrence or negative overlap) versus those that do not account for *d*. When the number of available attributes is finite and limited, those measures that account for *d* are more representative, as the non-occurrence of an attribute can be considered important when it is one of few possible attributes; when the number of attributes is large (including infinite) the measures that do not account for *d* should be used as the co-non-occurrence of one of an infinite number of possible attributes is of little meaning.

k Vector Definitions

To expand the similarity measures discussed it is necessary to establish a set of *k* vectors (X_i for i=1 to k), within each of these vectors X_i is described by *n* attributes ($X_{i,j}$ is the attribute value for vector *i* for attribute *j*). Adding additional columns to Table 1 for additional X variables creates some problems when defining *b* and *c*. It remains easy to define *a* as all instances of an attribute being present in all Xs and *d* as all instances of an attribute being not present in all Xs. More Xs creates more options in the middle that do not cleanly fall into the definition of either *b* or *c* from the 2 variable formulation.

	X_1	X_2	X_3
a	1	1	1
?	1	1	0
?	1	0	1
?	1	0	0
?	0	1	0
?	0	1	1
?	0	0	1
d	0	0	0

Table 4: Three Variables

First, *c* is abandoned in favor of multiple formulations of *b* and rewrite Table 4 as follows:

	X_1	X_2	X_3
a	1	1	1
b_1	1	1	0
b_2	1	0	1
b_3	1	0	0
b_4	0	1	0
b_5	0	1	1
b_6	0	0	1
d	0	0	0

Table 4a: Three Variables with *b*'s

This resolves the limitations of the *b/c* nomenclature but creates extensive formulations. Upon examination of the many similarity measures in existence it becomes apparent that most use some function of p_1 and/or p_2 in the denominator. From Table 3 we can see that p_1 is simply the sum of X_1 over all attributes, from that we expand the *p* measures to include one for each column and the *q* measures simply become the complement of the *p*'s, i.e. $q_1=n-p_1$. This formulation allows for multi-variable approaches to existing measures such as (for three variables):

$$S_{Dice-3} = \frac{3a}{p_1 + p_2 + p_3}$$

This works because the denominator for Dice (in the two variable format) is made up of *(a+b)+(a+c)* or p_1+p_2. Dice (and other similar measures) lends itself nicely to expansion in this form and results in (for *k* variables):

$$S_{Dice-k} = \frac{ka}{p_1 + p_2 + \cdots + p_k}$$

and

$$S_{Sorgenfrei-k} = \frac{a^k}{p_1 * p_2 * \ldots * p_k}$$

In the case of similarity measures that do not use such a formulation like Jaccard, some other form can be used, e.g. $S_{Jaccard} = \frac{a}{n-d}$ where *n* remains the number of attributes; in fact this formulation of Jaccard works for any number of vectors. The simple matching measure even retains its original form: $S_{Sokal-Michener} = \frac{a+d}{n} = a' + d'$.

These multi-variable similarity measures yield the similarity between all vectors; as such it is advisable to only compare three way measures with three way measures when evaluating for which group is more similar than another, as three-way measures will always be smaller than two-way measures, etc. of similar vectors (as $p_1+p_2+\ldots+p_k$ goes up S goes down).

Non-binary Measures

Until now the vast majority of the proposed measures of similarity have been based on the presence or absence of binary attributes. There exist, however, countless instances of attributes being present within an individual that are not expressed completely. These cases

represent a problem for traditional similarity measures; should the partial presence be treated as presence or absence, or is there some third option?

Conceptual Formulation

Consider two vectors, X_1 and X_2, each described by n attributes; each attribute can be expressed as a percentage at which it is expressed in the vector, e.g. X_1 expresses 27% of attribute 1. The sum of the strengths of the individual attributes within each vector do not necessarily add up to 100%, nor do the sum of strengths across vectors add to 100%. This type of test data was used to evaluate potential measures of similarity and distance.

Data generation and testing. First, a random data set was created to assess various trial similarity measures. Two vectors of n=100 were generated in Excel; the first vector using rand() and the second vector as a function of the first vector. A weight of between 0.7 and 1.0 was applied to the first vector randomly to create the second vector such that it is no more than 30% less in all values than the first. Since this weighting was via a uniformly distributed random number (Appendix B), it is expected that X_2 will be 15% less than X_1 on average. Therefore should yield a similarity approximately equal to 0.85 by city-block distance.

The first attempt at a percentage similarity measure began with calculating the Pearson Product Moment for the two vectors. This test yielded a result of r=0.983. The Pearson correlation breaks down in cases of one vector attribute value equaling zero with the other vector having some non-zero value (*b* or *c* cases from binary similarity). Pearson reduces the numerator for each *b* or *c*, however since the other vector may be very near zero there is no reason why close proximity (e.g. between 0 and 0.1) should be treated differently than between 0.8 and 0.9.

This indicates that a similarity measure that does not separate out *b* and *c* cases but calculates distance regardless of the values is required.

Negative overlap. The question of how to approach negative overlap, the condition *d* in the binary measures should be answered before getting too deep into evaluating proposed measures. In cases where the co-non-occurrence (negative overlap) of an attribute is important to the overall picture, then *d* must be accounted for. This situation is likely to occur when there are a small finite number of attributes. If two vectors each have ten attributes and four of them are negative overlaps it indicates that within those four descriptive areas the two vectors are the same. If two vectors have the potential to include infinite attributes but only ten are reported, those four negative overlaps are probably a very small percentage of total negative overlaps, and are much less likely to be indicative of an overall similarity. Going forward this decision rule will be applied; if the number of attributes is small and from a finite population, then *d* will be included, if the number of attributes is large and/or from an infinite (or very large) population, then *d* will be discounted. Most formulations throughout the remainder of this paper will specifically address *d* by removing it from the similarity (*a*) value (distance based measures often result in calculating *a+d*, so *d* needs to be subtracted when it adds no value).

Further testing of proposed measures. The second attempt at non-binary similarity involves a percentage similarity. A percentage similarity $\frac{X_{j,max}-X_{j,min}}{X_{j,max}}$ *or* $\frac{|X_{1,j}-X_{2,j}|}{X_{1,j}}$ will vary depending upon where the values lie between 0 and 1. Previously it was discussed that the difference between 0.9 and 0.8 should be treated with equal weight to the difference between 0.1 and 0.2. The closer to 1 the more weight the relationship will be given, as such this method can be eliminated from consideration (although in the case of this specific data it does come close to

our target of 85% at 85.58%, however this is more a function of the data specification than the quality of the method). A distance measure does not weight differences at the 1 end of the spectrum more. At their most basic, a measure of distance is simply the difference between the two values. By applying some treatment to that value, either absolute value, or squaring, a positive value can be arrived at regardless of the larger value. This method (when the absolute value is taken) would yield 0.1 in both of the scenarios above, thus treating them as the same difference.

Advancing the absolute value method to calculate the arithmetic mean over *n* attributes yields: $D' = \sum \frac{|X_1 - X_2|}{n}$ (with similarity being simply the complement, *S=1-D'*). This is the arithmetic average city-block or Manhattan distance between the two vectors and will be seen more in this paper. Other distance measures can also be used here as well. Euclidean Distance can be used as it treats the differences similarly (although squared instead of as an absolute value) giving: $D = \sqrt{\sum (X_1 - X_2)^2}$ the arithmetic average of which is: $D' = \frac{\sqrt{\sum (X_1 - X_2)^2}}{n}$.

Interestingly, the common formulation of $D' = \sqrt{\frac{\sum (X_1 - X_2)^2}{n}}$ represents the quadratic mean of the Euclidean distances (as opposed to the arithmetic mean used above) this value is the same for the quadratic mean of the city-block distance $D' = \sqrt{\frac{\sum |X_1 - X_2|^2}{n}}$. A choice to use the quadratic mean over the arithmetic mean eliminates the choice between city-block and Euclidean distance.

Minkowsky distance is a generalization of Euclidean distance where $D = \sqrt[r]{\sum |X_1 - X_2|^r}$. Minkowsky distance forms the Euclidean distance for *r*=2, the city-block distance for *r*=1 and the Chebychev distance for *r*=∞ (for Chebychev distance, $D = \sum \max |X_1 - X_2|$). Additional

distance measures, such as Chi-square, Mahalanobis, Quadratic, or Canberra can also be used. Mahalanobis distance ($d = \sqrt{(X_{1,j} - X_{2,j})S^{-1}(X_{1,j} - X_{2,j})^T}$ tends to be very resource intensive to calculate for large data sets. Chord distance, which is a transformation of Cosine distance can overcome the problems in non-normalized Euclidean distance. Without percentage normalization, Euclidean distance is sensitive to outliers (large values of attributes mask smaller values), and even with normalization, if attributes repeat they will be more heavily weighted.

For the purposes of this analysis, attribute values are percentage normalized, thus eliminating the outlier issue with Euclidean distance (the same issue arises in Minkowsky distance for all values of *r*). To normalize individual values: for attribute *j* the value attributed to X_1 is equal to the raw value minus the minimum divided by the range for that attribute (the range being equal to the max possible value if the minimum is zero) thus giving: $\hat{X}_{1,j} = \frac{X_{1,j} - X_{1,j,min}}{X_{1,j,max} - X_{1,j,min}}$. For situations where the minimum is zero, X_{min} can obviously be removed. Alternately the value can be divided by the standard deviation of the attribute instead of its range to trim outliers, this may require mapping values that exceed either end (0,1) to the limits (Wilson & Martinez 1997) (this method will not be explored in this analysis). By treating all attributes as percentages, it eliminates the weight problem discussed, provides a standard reference for all attribute values (0,1), and allows for selective weighting of attributes down the road.

k vector percentage normalized metric

To turn the above discussed distances into a metric measure that can be treated as the complement of similarity a decision point needs to be addressed first. Since the states of overlap (positive and negative) can still exist in a non-binary situation, they need to be addressed as either both cases of zero distance, or as something else. Building from the Heterogeneous

Euclidean-Overlap Metric (HEOM) (Wilson & Martinez 1997) and replacing Euclidean distance with Minkowsy distance (allowing for adjustment in *r* to yield several various distance measures) provides a distance measure to start from. Since the distance between attributes is important in all cases (and not just the case of complete positive overlap) the *b* and *c* measures are moot. Distance will yield 0 for both the previously defined *a* or *d* case.

Positive overlap (*a*) can be ignored as all cases that are not negative overlap are some form of dissimilarity that can be measured by the distance between vectors (*a* having zero distance). Cases of all vectors having some value are fundamentally no different from cases where one vector has a value and the remaining *k*-1 vectors are all zero; there is still a distance between them that can be measured and represents their dissimilarity.

Negative Overlap can be ignored when a small finite number of attributes are in question. If the number of possible attributes is small and finite the situation of negative overlap indicates that the vectors are similar in the non-presence of that attribute. If the number of possible attributes is large (or infinite) negative overlap may not be important in understanding similarity (or dissimilarity) between the vectors. If for example the degree to which *k* vectors represent the attribute “Tastes like Chicken” is included and the vectors each represent a planet in the solar system, then the negative overlap tells us nothing and should be discarded. If a large number is in question then the decision to discard negative overlaps should be made. This case is the *d* measure from binary similarity which can be calculated in R by *d*=(Sum(Trunc((1-X_1)*…*(1-X_k)), na.rm=TRUE)) this returns the number of cases where all X=0. This value can be subtracted from the numerator in the distance measure utilized to discount those cases of negative overlap resulting in $a=n-(D+d)$, using city-block distance (*D).*

Conveniently this formulation works whether the variables are binary or percentage normalized. This reduces HEOM to a single decision from the two it started with. That decision, whether to specifically discount negative overlap, remains.

***k*-vector distance.** City-block distance for a three vector formulation reduces to max minus min. Assuming that the first vector contains the max value, the second vector a value in the middle, and the third vector the minimum value the distance becomes the sum of the individual distances between 1 and 2 and 2 and 3, or: $(X_1-X_2)+(X_2-X_3)=X_1-X_3$. Since the city-block formulation includes the absolute value of those differences, the location of the min, max, and middle values becomes irrelevant, and the three vector version reduces to |max-min|, or simply max-min. From there it is an easy leap to a *k*-vector formulation, since the end result is the same. City-block benefits from simplicity in this case, and for that reason alone should be considered when evaluating a best distance measure for any *k*-vector data set.

The quadratic mean of city-block was previously shown to equal the quadratic mean of Euclidean distance, that holds true here as well. The arithmetic mean is equal to the quadratic mean times $n^{0.5}$, so by multiplying the quadratic mean of the *k*-vector city block distance by $n^{0.5}$ the result is the arithmetic mean of the Euclidean distance for *k*-vectors. This is a round-about way of getting there, but a useful tool to understand the interrelation between the two distance measures and the two means.

To equate city block distance back to 2-vectors, the distance calculated is equal to $n-(a+d)$, where *a* and *d* are as defined before. Since city-block equals zero whenever the two vectors have the same value, it equals zero for both positive and negative overlap (positive and negative co-occurrence). Sokal-Michener in *k*-vector becomes $(a+d)/n$, and is the arithmetic

mean of *n-D* when *D* is calculated via city-block distance, therefore the multi-vector percentage normalized version retains the same form with $a + d = \sum(1 - (max - min))$. The only need for a decision point in the modified HOEM is for the case of all X=0 (negative overlap), and that is only required when *n* is large or in specific cases of small *n* where the co-non-occurrence (negative overlap) doesn't yield valuable information.

This max-min formulation does not account for variation (beside the difference between max and min) within each attribute *j*. The variation is hidden within the variation between the outliers on either end, as such, it is not an accurate way of measuring the distance among the vectors, instead acting as a distance between tails.

By using standard deviation for a population within each attribute *j* we can account for all of the variance present between all Xs. Using population instead of sample because; although there may be more Xs in the universe, only the ones being evaluated are being evaluated, and thus are assumed to be the only ones in existence or at least the only ones that matter. Since this eliminates Bessel's correction it force normalizes the range to (0,1). This makes $a + d = n - \sum_{j=1}^{n} 2 * \sqrt{\frac{\sum_{i=1}^{k}(X_{i,j} - \overline{X_i})^2}{k}}$ using arithmetic average city block distance. The 2 normalizes to a max value of 1 since $max\sqrt{\frac{\sum_{i=1}^{k}(X_{i,j} - \overline{X_i})^2}{k}} = .5$ for X≤1 and *d* remains $d = \sum(trunc(\prod_{i=1}^{k}(1 - X_i))$ just as before. Utilizing the Minkowsky formulation for distance as the method for averaging the standard deviation derived distance produces an equation that is customizable in results through modification of the exponent term (*r*), thus allowing a single equation to yield arithmetic, quadratic, and geometric (among other) averages.

$$a' + d' = 1 - \sqrt[r]{\left(\frac{\sum_{j=1}^{n} 2 * \sqrt{\frac{\sum_{i=1}^{k}(X_{i,j} - \overline{X_i})^2}{k}}}{n}\right)^r}$$

This Minkowsky version of the arithmetic average of the sum of standard deviations includes all of the variance between vectors over each attribute, is limited to (0,1), and meets the metric rules described by Zhang and Srihari (2003). In all, it appears to be the best approach to *k*-vector percentage normalized similarity, as it solves the problems noted with the various other attempted similarity measures. Formulations for *a, d, p_1, p_2*, etc. can be found in Appendix C.

Application

The most readily apparent application for the above distance measure is in determining similarity between *k* non-binary vectors (through four steps). In the first step percentage normalize the data as described above. In the second step determine *r* (*r*=2 for Euclidean distance, *r*=1 for city-block distance, and *r*=∞ for Chebychev distance). In the third step determine if negative overlap (*d*) is significant for the similarity being evaluated. In the fourth step, for a generalized version of $S_{\text{Russell\&Rao}}$ apply:

$$S_{R\&R\ Universal} = a' = 1 - \sqrt[r]{\left(\frac{\sum_{j=1}^{n} 2 * \sqrt{\frac{\sum_{i=1}^{k}(X_{i,j} - \overline{X_i})^2}{k}}}{n}\right)^r} - \sqrt[r]{\left(\frac{\sum_{j=1}^{n}(trunc(\prod_{i=1}^{k}(1 - X_i)))}{n}\right)^r}.$$

Alternately, for traditional measures, calculate

$a = n - \sum_{j=1}^{n} 2 * \sqrt{\frac{\sum_{i=1}^{k}(X_{i,j} - \overline{X_i})^2}{k}} - \sum_{j=1}^{n}(trunc(\prod_{i=1}^{k}(1 - X_i)))$ and all *k* values for $p_i = \sum_{j=1}^{n} 1 - (trunc(1 - X_{i,j})$, using those values Dice, Jaccard, etc. can be readily calculated.

Some measures use $a+b+c$ as the denominator, in those cases either use n-d or $(\sum_{i=1}^{k} p_i) - a(k-1)$ with the first method producing Jaccard values closer to Dice and the later closer to Sorgenfrei.

Egghe (2010) suggests that a good test of the quality of a similarity measure is whether the addition of a constant attribute value to both vectors results in an increase in similarity. This test was applied to the $S_{R\&R\ Universal}$ measure calculated with the data in Appendix B (two vector) by adding an additional attribute with a value of 0.5 to each vector. This addition cause the resultant similarity value to increase from 0.93067 to 0.93136, passing the Egghe test.

For prediction; utilizing *S* as a probability of inclusion from X_1 to X_2 for items or attributes exhibited by X_1 but not X_2 may prove useful. Used in this manner *S* becomes the slope for the new attribute in what amounts to a Probit model (DeSarbo, et al. 1987), such that the strength of the attribute in X_1 times *S* yields a value for the predicted strength of the same attribute in X_2.

Similarity measures are typically used to measure a percentage relationship between two or more vectors, however these results do not yield a significance component and are left up to the researcher to decide whether a relationship exists beyond chance and whether that relationship is of value. Since traditional similarity measures are nonparametric in nature, there are no distinct criteria for significance and no accepted statistical tests. Establishing arbitrary cutoffs (50%) allow for some form of hypothesis testing, however care must be taken to define the rejection criteria before the test is performed so as not to influence the results (*p* hacking).

Conclusion

The nomenclature utilized for binary similarity measures has been largely accepted within the field, however there has still remained some aspects that were not clear. The use of a' to indicate the average of a over n resolves the problem of ambiguity. Although some authors use set or Boolean notation to indicate the different agreement conditions, a standardized arithmetic notation eliminates confusion across disciplines.

Generalizing binary measures of similarity from 2 vectors to k vectors has proven to be of relative ease once the conditions of b and c are addressed. By ignoring them and treating them as the remainder once a and d are determined, or by utilizing measures that do not use them specifically but rather rely on p_1 through p_k instead, the problem of their definitions becomes moot.

Non-binary attribute values present several problems, most of which are addressed through percentage normalization. Once normalized there are several distance measures to choose from. The most universal and flexible method involves Minkowsky distance, allowing for determination of r as a method of moving between city-block and Euclidean distances. This formulation is further expandable in the means of averaging over n. An extension of the Minkowsky distance formula allows for the average over n to be arithmetic, quadratic, etc. proving useful when evaluating amongst the options.

Expanding the non-binary measures to k vectors results in the decision to abandon max-min in favor of standard deviation, which has been shown to be equivalent to the arithmetic average of the Euclidean distance between each vector i within each attribute j and the average of same. For a two vector model this reduces to a max-min formulation, but for k vectors it includes

the variability within the attribute (which is not included in max-min), providing a better picture of overall within-attribute distance.

Using the standard deviation model a test for significance has been developed and shown to produce results consistent with expectations for randomly generated data and for modified randomly generated data. When treated as a Probit model as suggested by DeSarbo, et al. (1987), this new model can help improve understanding with regard to how individuals and groups are interrelated in numerous fields.

References

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

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian journal of ecology*, *18*(1), 117-143.

Consonni, V., & Todeschini, R. (2012). New similarity coefficients for binary data. *Match-Communications in Mathematical and Computer Chemistry*, *68*(2), 581.

De Benedictis, L., & Tajoli, L. (2007). Economic integration and similarity in trade structures. *Empirica*, *34*(2), 117-137.

DeSarbo, W. S., De Soete, G., & Eliashberg, J. (1987). A new stochastic multidimensional unfolding model for the investigation of paired comparison consumer preference/choice data. *Journal of Economic Psychology*, *8*(3), 357-384.

Egghe, L. (2010). Good properties of similarity measures and their complementarity. *Journal of the American Society for Information Science and Technology*, *61*(10), 2151-2160.

Eric, O. N., Gilbert, J. F., Marshall, K. G., & Oladi, R. (2015). *A New Measure of Economic Distance* (No. 5362). CESifo Group Munich.

Jones, D. L., PhD. (2016). Analysis of Similarity (ANOSIM). Retrieved June 03, 2016, from http://www.marine.usf.edu/user/djones/anosim/anosim.html

Kim, J., & Billard, L. (2013). Dissimilarity measures for histogram-valued observations. *Communications in Statistics-Theory and Methods*, *42*(2), 283-303.

Lourenço, F., Lobo, V., & Bação, F. (2006) Binary-based similarity measures for categorical data and their application in Self-Organizing Maps. Retrieved June 08, 2016, from http://www.ai.rug.nl/nl/vakinformatie/sr/articles/categorical-distances.pdf

Marks, R. E. (2013). Validation and model selection: Three similarity measures compared. *Complexity Economics*, *2*(1), 41-61.

Novak, Z., & Pap, Z. (2012). Exploiting interest-based proximity for content recommendation in peer-to-peer networks. *Communications, IET*, *6*(12), 1595-1601.

Shirkhorshidi, A. S., Aghabozorgi, S., & Wah, T. Y. (2015). A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data. *PloS one*, *10*(12), e0144059.

Stein, B., Niggemann, O., & Husemeyer, U. (2000). Learning Complex Similarity Measures. In *Classification and Information Processing at the Turn of the Millennium* (pp. 254-263). Springer Berlin Heidelberg.

Stein, B., & Niggemann, O. (2001). Generation of similarity measures from different sources. In *Engineering of Intelligent Systems* (pp. 197-206). Springer Berlin Heidelberg.

Warrens, M. J. (2008). *Similarity coefficients for binary data: properties of coefficients, coefficient matrices, multi-way metrics and multivariate coefficients*. Psychometrics and Research Methodology Group, Leiden University Institute for Psychological Research, Faculty of Social Sciences, Leiden University.

Wilson, D. R., & Martinez, T. R. (1997). Improved heterogeneous distance functions. *Journal of artificial intelligence research*, 1-34.

Zhang, B., & Srihari, S. N. (2003). Properties of binary vector dissimilarity measures. In *Proc. JCIS Int'l Conf. Computer Vision, Pattern Recognition, and Image Processing* (Vol. 1).

Appendix A: Common Similarity Measures

Measure	2-variable	k-variable	% Attribute, k-variable
$S_{Russell\&Rao}$ (r=1)	$\frac{a}{n}$	$\frac{a}{n}$	$1-\sqrt[r]{(\frac{\sum_{j=1}^{n} 2 * \sqrt{\frac{\sum_{i=1}^{k}(X_{i,j}-\overline{X_i})^2}{k}}}{n})^r} - \sqrt[r]{(\frac{\sum_{j=1}^{n}(trunc(\prod_{i=1}^{k}(1-X_i)))}{n})^r}$
$S_{SM} = S_{city-block}$ (r=1)	$\frac{a+d}{n}$	$\frac{a+d}{n}$	$1-\sqrt[r]{(\frac{\sum_{j=1}^{n} 2 * \sqrt{\frac{\sum_{i=1}^{k}(X_{i,j}-\overline{X_i})^2}{k}}}{n})^r}$
S_{Dice}	$\frac{2a}{p_1+p_2}$	$\frac{ka}{p_1+p_2+\cdots+p_k}$	$\frac{k(n-\sum_{j=1}^{n} 2 * \sqrt{\frac{\sum_{i=1}^{k}(X_{i,j}-\overline{X_i})^2}{k}} - \sum_{j=1}^{n}\left(trunc(\prod_{i=1}^{k}(1-X_i))\right))}{\sum_{i=1}^{k} p_i}$
$S_{Sorgenfrei}$	$\frac{a^2}{p_1 * p_2}$	$\frac{a^k}{p_1 * p_2 * \ldots * p_k}$	$\frac{(n-\sum_{j=1}^{n} 2 * \sqrt{\frac{\sum_{i=1}^{k}(X_{i,j}-\overline{X_i})^2}{k}} - \sum_{j=1}^{n}\left(trunc(\prod_{i=1}^{k}(1-X_i))\right))^k}{\prod_{i=1}^{k} p_i}$
$S_{Jaccard}$	$\frac{a}{a+b+c}$	$\frac{a}{n-d}$ or $\frac{a}{(\prod_{i=1}^{k} p_i)-a(k-1)}$	$\frac{(n-\sum_{j=1}^{n} 2 * \sqrt{\frac{\sum_{i=1}^{k}(X_{i,j}-\overline{X_i})^2}{k}} - \sum_{j=1}^{n}\left(trunc(\prod_{i=1}^{k}(1-X_i))\right))}{n-\sum_{j=1}^{n}\left(trunc(\prod_{i=1}^{k}(1-X_i))\right)}$

Appendix B: Development Data

X_1	X_2
0.340602	0.275888
0.68802	0.639858
0.821848	0.821848
0.397286	0.345639
0.263364	0.229127
0.035587	0.035587
0.157219	0.119486
0.213186	0.153494
0.12093	0.111255
0.284669	0.216349
0.288161	0.247818
0.05854	0.057369
0.093149	0.081971
0.071666	0.063066
0.525647	0.37321
0.738247	0.671805
0.208882	0.154573
0.942096	0.923254
0.183399	0.172395
0.548421	0.449705
0.280547	0.238465
0.67324	0.498198
0.597845	0.478276
0.602816	0.548562
0.711556	0.661747
0.968161	0.958479
0.564296	0.49658
0.779835	0.701852
0.296534	0.278742
0.25166	0.186229
0.945287	0.860211
0.71601	0.529848
0.597892	0.538103
0.544292	0.386447
0.182031	0.147445
0.417878	0.338481
0.916226	0.659683
0.715954	0.701635

X_1	X_2
0.306361	0.242026
0.994146	0.845024
0.208921	0.200564
0.985705	0.887134
0.370664	0.329891
0.173636	0.145854
0.435506	0.309209
0.367051	0.308323
0.555638	0.488961
0.785819	0.565789
0.112036	0.110916
0.832574	0.632756
0.12971	0.123225
0.697062	0.648267
0.068136	0.067455
0.785189	0.565336
0.863307	0.699279
0.929838	0.855451
0.400129	0.288093
0.957891	0.919576
0.690108	0.593493
0.450693	0.324499
0.981283	0.922406
0.42387	0.38996
0.044073	0.034377
0.059858	0.051478
0.748699	0.658855
0.61662	0.591955
0.660463	0.647253
0.231882	0.176231
0.788843	0.733624
0.208004	0.158083
0.277543	0.269216
0.082125	0.073091
0.676292	0.64924
0.922073	0.857528
0.936786	0.843107
0.433612	0.403259

X_1	X_2
0.147197	0.108926
0.915749	0.714284
0.158594	0.120531
0.195504	0.138808
0.89086	0.739414
0.714764	0.557516
0.569555	0.438558
0.624775	0.449838
0.025385	0.0231
0.343074	0.298474
0.58016	0.551152
0.249354	0.249354
0.823335	0.691601
0.77277	0.57185
0.706323	0.614501
0.427081	0.345936
0.525702	0.473132
0.94578	0.784997
0.316947	0.250388
0.178148	0.142518
0.244469	0.232245
0.965336	0.888109
0.080829	0.063855
0.060583	0.052102

Pearson $r=0.983238$

Percentage Similarity $D'=0.1442,\ S=0.8558$

Quadratic Average Euclidean Distance $D'=0.0918,\ S=0.9082$

$S_{R\&R\ Universal}\ S=0.9307$

$S_{Jaccard}\ S=0.9307$

Modifying the data to replace the last set with 0,0 (negative overlap) affects R&R and Jaccard as follows:

$S_{R\&R\ Universal}\ S=0.9208$

$S_{Jaccard}\ S=0.9301$

Jaccard does not decrease as much, because, although in both cases the numerator decreases, in the case of Jaccard, so does the denominator.

Appendix C: k-vector Percentage Normalized Values

$$(a+d) = n - \sum_{j=1}^{n} 2 * \sqrt{\frac{\sum_{i=1}^{k}(X_{i,j} - \overline{X_{\iota}})^2}{k}}$$

$$d = \sum_{j=1}^{n} (trunc\left(\prod_{i=1}^{k}(1 - X_i)\right))$$

$$a = n - \sum_{j=1}^{n} 2 * \sqrt{\frac{\sum_{i=1}^{k}(X_{i,j} - \overline{X_{\iota}})^2}{k}} - \sum_{j=1}^{n} (trunc\left(\prod_{i=1}^{k}(1 - X_i)\right))$$

$$p_i = \sum_{j=1}^{n} 1 - trunc(1 - X_{i,j})$$

$a + b + c = (\sum_{i=1}^{k} p_i) - a(k-1)$ or $n - \sum_{j=1}^{n}(trunc(\prod_{i=1}^{k}(1 - X_i)))$

Acknowledgements

I would like to thank Emil Berendt (SNHU) and Scott Barrow for providing formatting, grammar, and usage reviews. Emil again for also reviewing for readability and understanding from an Economist's prospective. I would like to especially thank Jeff Picka (UNB) for helping me to resolve some statistical questions related to significance testing and the use of F tests in general.