Binary Distances Through Similarity-Dissimilarity Measures

Should we want to compare two objects U and V as far as their similarity is concerned, then in the binary case there are only four possible outcomes, which are presented in the table below:

		Object U		
	Outcome	1	0	Total
Object V	1	a	b	$a+b$
	0	c	d	$c+d$
	Total	$a+c$	$b+d$	$a+b+c+d$

Table 1: A 2×2 Contingency table between two partitions (objects) U and V.

What we can say about these two objects is how similar or dissimilar they are. If they share the same number of ones and zeroes (or a and d in the table above), then they can be considered similar to each other, if however they do not, then these two objects are characterized as dissimilar. Here are some similarity measures that use the table above (*Seung-Seok Choi, Sung-Hyuk Cha, Charles C. Tappert, 2009*):

Similarity-Dissimilarity Measures	
Similarity Measure	**Formula**
Jaccard or Tanimoto	$\frac{a}{a+b+c}$
Sørensen–Dice or Czekanowski or Nei & Li	$\frac{2a}{2a+b+c}$
3W - Jaccard	$\frac{3a}{3a+b+c}$
Simple Matching or Sokal & Michener	$\frac{a+d}{a+b+c+d}$
Sokal & Sneath I	$\frac{a}{a+2(b+c)}$
Sokal & Sneath II	$\frac{2(a+d)}{2a+b+c+2d}$
Sokal & Sneath III	$\frac{a+d}{b+c}$

Sokal & Sneath IV	$\frac{1}{4}\left(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{b+d}+\frac{d}{c+d}\right)$
Sokal & Sneath V or Ochiai II	$\frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$
Rogers & Tanimoto	$\frac{a+d}{a+2(b+c)+d}$
Faith	$\frac{a+0.5d}{a+b+c+d}$
Gower & Legendre	$\frac{a+d}{a+0.5(b+c)+d}$
Russell & Rao	$\frac{a}{a+b+c+d}$
Ochiai I or Cosine or Otsuka-Ochiai or Fowlkes-Mallows Index	$\frac{a}{\sqrt{(a+b)(a+c)}}$
Forbes I	$\frac{na}{(a+b)(a+c)}$
Forbes II	$\frac{na-(a+b)(a+c)}{n[min(a+b,a+c)-(a+b)(a+c)]}$
Fossum	$\frac{n(a-0.5)^2}{(a+b)(a+c)}$
Sorgenfrei	$\frac{a^2}{(a+b)(a+c)}$
Mountford	$\frac{a}{0.5(ab+ac)+bc}$
McConnaughey	$\frac{a^2-bc}{(a+b)(a+c)}$

Tarwid	$\frac{na-(a+b)(a+c)}{na+(a+b)(a+c)}$
Kulczynski I	$\frac{a}{b+c}$
Kulczynski II or Driver & Kroeber	$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$
Johnson	$\frac{a}{a+b}+\frac{a}{a+c}$
Dennis	$\frac{ad-bc}{\sqrt{n(a+b)(a+c)}}$
Simpson or Szymkiewicz-Simpson	$\frac{a}{min(a+b,a+c)}$
Braun-Blanquet	$\frac{a}{max(a+b,a+c)}$
Fager & McGowan	$\frac{a}{\sqrt{(a+b)(a+c)}}-\frac{max(a+b,a+c)}{2}$
Gower	$\frac{a+d}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$
Pearson I	$\frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$
Pearson II	$\sqrt{\frac{\frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}}{n+\frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}}}$

Pearson III	$\sqrt{\frac{\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}}{n+\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}}}$
Pearson & Heron I	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$
Pearson & Heron II	$cos\left(\frac{\pi\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}\right)$
Cole	$\frac{\sqrt{2}(ad-bc)}{\sqrt{(ad-bc)^2-(a+b)(a+c)(b+d)(c+d)}}$
Stiles	$log_{10}\left(\frac{n\left(\lvert ad-bc\rvert-{}^{n}/_{2}\right)^2}{(a+b)(a+c)(b+d)(c+d)}\right)$
Yule's Q	$\frac{ad-bc}{ad+bc}$
Yule's ω or Yule's Y	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$
Dispersion	$\frac{ad-bc}{(a+b+c+d)^2}$
Hamann	$\frac{(a+d)-(b+c)}{a+b+c+d}$
McEwen & Michael	$\frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$
Goodman & Kruskal's Lambda	$\frac{\sigma-\sigma'}{2n-\sigma'}$ with: $\sigma = max(a,b)+max(a,c)+max(b,d)+max(c,d)$ $\sigma' = max(a+c,b+d)+max(a+b,c+d)$

Anderberg	$\frac{\sigma - \sigma'}{2n}$ with: $\sigma = max(a,b) + max(a,c) + max(b,d) + max(c,d)$ $\sigma' = max(a+c, b+d) + max(a+b, c+d)$
Baroni-Urbani & Buser I	$\frac{\sqrt{ad} + a}{\sqrt{ad} + a + b + c}$
Baroni-Urbani & Buser II	$\frac{\sqrt{ad} + a - (b + c)}{\sqrt{ad} + a + b + c}$
Peirce	$\frac{ad - bc}{(a+b)(c+d)}$
Eyraud	$\frac{a - (a+b)(a+c)}{(a+b)(a+c)(b+d)(c+d)}$
Tarantula	$\frac{\frac{a}{a+b}}{\frac{a}{a+b} + \frac{c}{c+d}}$

Table 2: A table that contains all the similarity measures that will be calculated, for binary data.

The distance is calculated in the following manner:

$$Distance = 1 - Similarity\ Measure$$

How one picks a suitable distance depends on how they want to treat the zero-zero similarity matches (or d in the table above), as well as the dissimilarities between the objects (meaning b and c in the table above).

REFERENCES

Choi, Seung-Seok, Sung-Hyuk Cha and Charles C. Tappert. "*A Survey of Binary Similarity and Distance Measures*". Journal on Systemics, Cybernetics and Informatics (8): 43-48, 2010.

Zdenek Hubálek. "*Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation*". Biological Reviews, 57(4):669–689, February 2008. doi:10.1111/j.1469-185X.1982.tb00376.x.

Charles S. Peirce. "*The numerical measure of the success of predictions. Science*", 4(93):453–454, 1884.
doi:10.1126/science.ns-4.93.453-a.

James A. Jones and Mary Jean Harrold. "*Empirical evaluation of the tarantula automatic fault-localization technique*". In ASE '05 Proceedings of the 20th IEEE/ACM international Conference on Automated software engineering, 273–282. New York, November 2005. ACM, ACM.
doi:10.1145/1101908.1101949.