Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2021;13(3):282-92

ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH DOI: 10.5336/biostatic.2021-84763

İkili Veriler İçin Benzerlik Katsayılarının Değerlendirilmesi: Bir Benzetim Çalışması

Evaluation of Similarity Coefficients of Binary Data: A Simulation Study

İsmet DOĞAN[a], Nurhan DOĞAN[a], Taylan DOĞAN[b]

[a]Afyonkarahisar Sağlık Bilimleri Üniversitesi Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim ABD, Afyonkarahisar, TÜRKİYE
[b]KPN B.V. Mobile Telecommunications Company, Amsterdam, HOLLANDA

ÖZET Amaç: Bu çalışmanın amacı, türetilmiş veri setleri kullanarak farklı n, a, b, c *ve* d değerleri için belirlenen 72 farklı ikili benzerlik katsayısını tanıtmak, özelliklerini ortaya koyarak değerlendirmektir. **Gereç ve Yöntemler:** Bu çalışmada, ikili veriler için ileri sürülen benzerlik katsayıları ele alınmıştır. Çalışmada Phyton-random kütüphanesi kullanılarak $10 \leq n \leq 1000$ aralığında yer alan 35 farklı n değeri için veri türetilmiştir. Verilerin türetilmesinde önce a, b, c *ve* d ile gösterilen gözelerden hangisine değer atanacağı sonra da ilgili gözeye atanacak değer belirlenmiştir. $n = 10$ için 288, $n = 15$ için 817 ve $n \geq 20$ için 1000'er farklı veri seti çalışmada kullanılmıştır. **Bulgular:** İkili veriler için tüm benzerlik katsayılarının değer aralığının 0 (benzerlik yok) ile 1 (tam benzerlik) olması beklenmesine rağmen tüm katsayılar için bu aralık geçerli değildir. Dikkate alınan 72 farklı katsayı içerisinden 29 tanesi bu aralıkta değer almaktadır. Hiyerarşik Kümeleme Analizi'ne göre benzerlik katsayılarının çoğu birbirine benzemektedir. **Sonuç:** Genel olarak hemen tüm katsayılara ait değerler, örnekler daha benzer hâle geldikçe sabit bir minimumdan sabit bir maksimuma doğru artmaktadır. Ancak Hamann ve Sokal-Michener tarafından önerilen katsayılar, tüm n değerleri için benzerlik ile doğrusal olarak sorunsuz bir şekilde artmaktadır. Değer aralığının 0-1 olması ve benzerlik artışı ile paralellik göstermesinden dolayı Sokal-Michener tarafından önerilen katsayı tüm katsayılar içerisinde öne çıkmaktadır. Eyraud, Fager-McGowan, Fossum, Gower, Harris-Lahey, Pearson I ve Sokal-Sneath IV benzerlik katsayıları n sayısından etkilenmekte diğer katsayılar etkilenmemektedir. Dolayısıyla benzerlik katsayılarının önemli bir kısmının örnek büyüklüğünden bağımsız oldukları belirlenmiştir.

Anahtar kelimeler: Benzerlik katsayısı; hiyerarşik kümeleme; ikili veri

ABSTRACT Objective: The aim of this study is to introduce 72 different binary similarity coefficients determined for different n, a, b, c *and* d values by using derived data sets and to evaluate them by revealing their properties. **Material and Methods:** In this study, the similarity coefficients put forward for binary data are considered. In the study, data were derived for 35 different n values in the range of $10 \leq n \leq 1000$ using the Phyton-random library. In the derivation of the data, firstly, which cell shown with a, b, c *and* d will be assigned value, then the value to be assigned to the relevant cell was determined. 288 for $n = 10$, 817 for $n = 15$ and 1000 different data sets for $n \geq 20$ were used in the study. **Results:** Although the value range of all similarity coefficients for binary data is expected to be 0 (no similarity) to 1 (exact similarity), this range is not valid for all coefficients. Out of 72 different coefficients, 29 take values in this range. According to the Hierarchical Clustering Analysis, most of the similarity coefficients are similar. **Conclusion:** In general, the values of almost all coefficients increase from a fixed minimum to a fixed maximum as the samples become more similar. However, the coefficients proposed by Hamann and Sokal-Michener increase smoothly linearly with similarity for all n values. The coefficient suggested by Sokal-Michener stands out among all coefficients because the value range is 0-1 and shows parallelism with the increase in similarity. Eyraud, Fager-McGowan, Fossum, Gower, Harris-Lahey, Pearson I and Sokal-Sneath IV similarity coefficients are affected by the number of n and other coefficients are not.

Keywords: Similarity coefficient; hierarchical clustering; binary data

Correspondence: Nurhan DOĞAN
Afyonkarahisar Sağlık Bilimleri Üniversitesi Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim ABD, Afyonkarahisar, TÜRKİYE/TURKEY
E-mail: nurhandogan@hotmail.com

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 31 May 2021 ***Received in revised form:*** 12 Jul 2021 ***Accepted:*** 13 Aug 2021 ***Available online:*** 09 Sep 2021

Benzerlik, incelenen özelliğe ait 2 farklı örnekten/değerlendiriciden (karar verici, cihaz vb.) elde edilen sonuçların niteliği (kalitesi, bileşimi vb.), özellikleri ve görünümü olarak ifade edilmekte, benzerlik katsayısı ise sonuçların aynı olma derecesi veya nesnelerin birbirine ne ölçüde benzediğini ölçmek için kullanılan bir işlev veya iki nesnenin benzerliğine karar vermek için yaygın olarak kullanılan ölçümler olarak tanımlanmaktadır.[1] İkili veriler için önerilen benzerlik katsayılarının özellikle taksonomik çalışmalarla ilgilenen bilim adamları tarafından ortaya konulan ilk tanımları 19. yüzyılın sonlarına kadar uzanmaktadır. Benzerlik katsayıları birçok uygulama alanında özellikle sınıflandırma, kümeleme, örüntü (pattern) vb. analizi ile ilgili problemlerde kritik rol oynamaktadır. Biyoinformatik, kemometri ve tıbbi alanların yanı sıra örüntü tanıma, veri madenciliği alanlarında birçok uygulamaya sahiptir.[2] Ekoloji ve biyocoğrafya ile ilgili araştırmalarda benzerlik katsayıları, özellikle türlerin bir arada varlığını ve örnekleme alanlarının benzerliğini incelemek için türler arası ilişki analizinde, kümeleme ve veri madenciliği uygulamalarında alt kümelere bir dizi gözlem veya veri atamak için kullanılmaktadır. Son zamanlarda iris ve parmak izi tanıma gibi tanımlama problemlerini çözmek için biyometrik alanlarda da uygulanmıştır. Sanal tarama için kullanılan benzerlik ölçüleri, moleküler parmak izlerinin kullanımına dayanmaktadır.[3] Performans, uygun bir ölçü seçimine dayandığından, birçok araştırmacı, yüz yılı aşkın bir süredir en anlamlı ikili benzerlik katsayısını bulmak için ayrıntılı çaba sarf etmiştir. Çeşitli alanlarda çok sayıda ikili benzerlik katsayısı önerilmiş olmasına rağmen bunların birlikte ele alınarak değerlendirildiği sadece birkaç karşılaştırmalı çalışma bulunmaktadır.[4-10] Benzerlik katsayıları, performans değerlendirmesini teşvik eden çok çeşitli alanlarda kullanılmaktadır. Örneklerin ikili durumunu evet/hayır, doğru/yanlış veya var/yok olarak ifade eden ikili özellikler, çok çeşitli verileri temsil etmek için yaygın olarak kullanılmıştır. İkili özellik matrisleri tarafından temsil edilen verileri değerlendirmek için uygun bir benzerlik katsayısının seçimi gereklidir, çünkü farklı benzerlik katsayıları çelişkili sonuçlar verebilmektedir.[9] Genel olarak, iki nesnenin bir dizi öznitelik bakımından evet/hayır, doğru/yanlış veya var/yok şeklinde karşılaştırılması söz konusu ise veya bir nesne kümesi, 2 sonuçlu 2 farklı değişken için çapraz sınıflandırılırsa 2×2 bir tablo elde edilir. Klasik bir 2×2 tablo örneği Tablo 1'de verilmiştir.

TABLO 1: Benzerlik katsayısı hesaplamaları için 2x2 tablo örneği.

		Değerlendirici Y		
		Evet	*Hayır*	*Toplam*
Değerlendirici X	*Evet*	a	b	$a+b$
	Hayır	c	d	$c+d$
	Toplam	$a+c$	$b+d$	$n=a+b+c+d$

Benzerlik katsayılarının hesaplanması $a, b, c\ ve\ d$ frekanslarına bağlıdır. $a, b, c\ ve\ d$ ile ifade edilen 4 frekans, X ve Y değerlendiricilerinin kararlarının ortak dağılımını karakterize eder. $a\ ve\ d$ frekansları sırasıyla pozitif ve negatif eşleşmeler olarak adlandırılır, $b\ ve\ c$ ise uyuşmazlıkları gösterir. Benzerlik katsayısı seçimi bazı kriterlere dayanmaktadır. Dikkate alınması gereken önemli bir nokta, hesaplamaya negatif eşlemeyi gösteren d'nin dâhil edilmesi veya hariç tutulmasıdır. Bazı veriler için her 2 nesnede de 1 öğenin olmaması benzerliği gösterir, ancak bazı durumlarda bu doğru olmayabilir. Bu nedenle benzerlik katsayıları genellikle 2 tipte kategorize edilir. İlk tür, negatif eşleşmeleri dikkate alır. İkinci türde ise hesaplama sırasında negatif eşleşmeler dikkate alınmamaktadır.[2] 2×2 tablolar için benzerlik katsayıları, 2 sonuçlu 2 farklı değişkenin ne ölçüde ilişkilendirildiğini veya 2 nesnenin birbirine ne ölçüde benzediğini ölçen fonksiyonlardır. Benzerlik katsayıları bağımsız değişken olarak $a, b, c\ ve\ d$ frekanslarını dikkate alan ve değişkenler arasındaki ilişki arttıkça daha yüksek sayısal değerlere döndüren işlevlerdir. Mevcut onlarca benzerlik katsayısı içerisinden hangi katsayının kullanılması gerektiği konusunda çok fazla kafa karışıklığı vardır. Benzerlik katsayıları, bazı istatistiksel parametrelerin tahmin edicileri değil, esas olarak tanımlayıcı katsayılar oldukları için özel katsayı türleridir. Çoğu benzerlik katsayısı için güvenilir güven aralıkları vermek zordur ve olası hatalar yalnızca bir tür randomizasyon işlemi ile tahmin edilebilir. İkili veriler için tüm benzerlik katsayıları-

nın aralığının 0 (benzerlik yok) ila 1,0 (tam benzerlik) olması beklenir. Ancak bu beklenti tüm katsayılar için geçerli değildir. Tüm uygulama alanları dikkate alındığında, benzerlik ölçülerinin taşıması beklenen 2 özelliği vardır. İlk olarak benzerlik ölçüsü, örnek büyüklüğünden ve topluluktaki türlerin sayısından bağımsız olmalıdır. İkinci olarak, 2 topluluk örneği daha benzer hâle geldikçe, ölçü sabit bir minimumdan sabit bir maksimuma sorunsuz bir şekilde artmalıdır.[1] İkili veriler için bugüne kadar literatürde yer alan, her biri kendi matematiksel özelliklerine sahip olan ve farklı bilimsel alanlarda kullanılan 72 farklı benzerlik katsayısı belirlenmiştir. Bazı katsayıların literatürde farklı isimlerle yer alması nedeniyle seçim sürecinde, aynı olan katsayıların (örneğin Jaccard I katsayısı Tanimoto katsayısı, Czekanowski katsayısı, Gleason katsayısı ya da Sørensen-Dice katsayısı olarak bilinir) hariç tutulmasına gayret edilmiştir. Çalışmada, ikili veriler için belirlenen 72 benzerlik katsayısı tanıtılmış ve birbirleriyle karşılaştırılmıştır. Bu çalışmanın amacı, türetilmiş veri setleri kullanarak farklı $n, a, b, c\ ve\ d$ değerleri için belirlenen 72 farklı ikili benzerlik katsayısını tanıtmak, özelliklerini ortaya koyarak değerlendirmektir. Makalede, Helsinki Deklerasyonu Prensipleri dikkate alınmıştır.

GEREÇ VE YÖNTEMLER

Literatürde yer alan benzerlik katsayıları sayısal ve yapısal veriler için kullanılan benzerlik katsayıları olmak üzere 2'ye ayrılmaktadır. Bu çalışmada, yapısal ve nicel veriler için ileri sürülen benzerlik katsayıları çalışma kapsamına dâhil edilmemiş olup özellikle ikili veriler için geliştirilmiş olan benzerlik katsayıları dikkate alınmıştır. İkili veriler için geliştirilmiş benzerlik katsayıları evet/hayır, doğru/yanlış veya var/yok biçiminde veriler mevcut olduğunda kullanılır ve bu nedenle nominal ölçek için uygundur. Bu çalışmada, özellikle ikili veriler için kullanılması önerilen benzerlik katsayıları ele alınmıştır, çünkü ikili veriler, çeşitli verileri temsil etmek için yaygın olarak kullanılmaktadır. Çalışmada, Phyton-random kütüphanesi kullanılarak $10 \leq n \leq 1000$ aralığında yer alan 35 farklı n değeri (10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 125, 150, 175, 200, 225, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1.000) için veri türetilmiştir. Verilerin türetilmesinde önce $a, b, c\ ve\ d$ ile gösterilen gözelerden hangisine değer atanacağı sonra da ilgili gözeye atanacak değer belirlenmiştir. $n = 10$ için 288, $n = 15$ için 817 ve $n \geq 20$ için 1000'er farklı veri seti çalışmada kullanılmıştır. $a, b, c, ve\ d$'ye dayalı 72 farklı benzerlik katsayısına ait formüller Tablo 2'de verilmiştir.

Ayrıca çalışmada, dikkate alınan benzerlik katsayılarının hangilerinin birbirine benzediğini belirlemek amacıyla hiyerarşik kümeleme yöntemi kullanılarak dendogramlar çizilmiştir. Kümelerin belirlenmesinde yöntem olarak Centroid Bağlantı yöntemi, uzaklık ölçüsü olarak ise Karesel Öklid Uzaklığı kullanılmıştır. Dendogramların elde edilmesinde SPSS 20.0 paket programından yararlanılmıştır.

BULGULAR

Bu çalışmada dikkate alınan benzerlik katsayılarının çoğu iyi bilinmekte ve yaygın olarak kullanılmaktadır. Herhangi bir model kullanmadan kapsamlı ve varsayımsal olarak rastgele elde edilen benzerlik matrislerine dayalı 72 farklı benzerlik katsayısına ait değerler elde edilmiş ve karşılaştırılmıştır. $n, a, b, c\ ve\ d$'nin farklı değerleri için elde edilen sonuçlar Tablo 3'te verilmiştir.

Tablo 3'te de görüldüğü üzere gözlem değerlerinin tamamının 2x2 tablodaki herhangi bir gözede yer alması durumunda katsayıların büyük bir kısmı hesaplanamamaktadır. Katsayıların büyük bir kısmı $a + d = n$ olduğunda maksimum, $b + c = n$ olduğunda ise minimum değerlerine ulaşmaktadır. Gözelerde yer alan gözlem değerlerinin eşit olması durumunda ise katsayılar çoğunlukla 0 ile 0,5 arasında değer almaktadır. Katsayıların hemen tamamı örnek büyüklüğünden etkilenmemektedir.

Otuz beş farklı n değeri için elde edilen dendogramlar sayfa sayısı fazlalığından dolayı ayrı ayrı çalışmada sunulmamıştır. Ancak dendogramlardan elde edilen sonuçlar Tablo 4'te verilmiştir.

TABLO 2: Benzerlik katsayıları.

No	Adı	Formül	Kronoloji	Değer aralığı
1	Peirce I[8,9]	$\frac{ad-bc}{(a+b)(c+d)}$	1884	$<-1,+1>$
2	Peirce II[5,8]	$\frac{ad-bc}{(a+c)(b+d)}$	1884	$<-1,+1>$
3	Doolittle[5]	$\frac{(ad-bc)^2}{[(a+b)(c+d)(a+c)(b+d)]}$	1885	$<0,+1>$
4	Yule I[5-7,9]	$\frac{ad-bc}{ad+bc}$	1900	$<-1,+1>$
5	Jaccard I[6,7,9]	$\frac{a}{a+b+c}$	1901	$<0,+1>$
6	Pearson I[5,7,9]	$\chi^2=\frac{n(ad-bc)^2}{[(a+b)(c+d)(a+c)(b+d)]}$	1905	$<0,+\infty)$
7	Pearson II[5,7,9]	$\sqrt{\frac{\chi^2}{(n+\chi^2)}}$	1905	$<0,+1)$
8	Forbes I[5-7,9]	$\frac{an}{(a+b)(a+c)}$	1907	$<0,+\infty)$
9	Jaccard II[7,9]	$\frac{3a}{3a+b+c}$	1912	$<0,+1>$
10	Yule II[5,7,9]	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$	1912	$<-1,+1>$
11	Yule III[5,6]	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(c+d)(b+d)}}$	1912	$<-1,+1>$
12	Czekanowski[6,7]	$\frac{2a}{2a+b+c}$	1913	$<0,+1>$
13	Pearson/ Hero[5,7,9]	$cos\left[\frac{180\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}\right]$	1913	$<-1,+1>$
14	Michael[5,7,9]	$\frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$	1920	$<-1,+1>$
15	Forbes II[5,7,9]	$\frac{[an-(a+b)(a+c)]}{[nmin\{(a+b),(a+c)\}-(a+b)(a+c)]}$	1925	$(-\infty,+1>$
16	Kulczynski I[5-7,9]	$\frac{1}{2}\left[\frac{a}{a+b}+\frac{a}{a+c}\right]$	1927	$<0,+1>$
17	Kulczynski II[5,6,9]	$\frac{a}{b+c}$	1927	$<0,+\infty)$
18	Braun/ Blanquet[5,7,9]	$\frac{a}{max\{(a+b),(a+c)\}}$	1932	$<0,+1>$
19	Driver/ Kroeber[5-7,9]	$\frac{a}{\sqrt{(a+b)(a+c)}}$	1932	$<0,+1>$
20	Eyraud[5,7,9]	$\frac{[a-(a+b)(a+c)]}{[(a+b)(c+d)(a+c)(b+d)]}$	1936	$<-1,0>$
21	Russell/ Rao[5-7,9]	$\frac{a}{n}$	1940	$<0,+1>$
22	Simpson[5-7,9]	$\frac{a}{min\{(a+b),(a+c)\}}$	1943	$<0,+1>$
23	Dice I[8]	$\frac{a}{(a+b)}$	1945	$<0,+1>$
24	Dice II[8]	$\frac{a}{(a+c)}$	1945	$<0,+1>$
25	Cole I[8]	$\frac{ad-bc}{(a+c)(c+d)}$	1949	$(-\infty,+1>$
26	Cole II[8]	$\frac{ad-bc}{(a+b)(b+d)}$	1949	$(-\infty,+1>$

TABLO 2: Benzerlik katsayıları *(devamı)*.

No	Adı	Formül	Kronoloji	Değer aralığı
27	Cole III[5,7,9]	$\frac{\sqrt{2}(ad-bc)}{\sqrt{[(ad-bc)^2+(a+b)(c+d)(a+c)(b+d)]}}$	1949	$<-1,+1>$
28	Goodman/Kruskal[7,9]	$\frac{2min(a,\ d)-b-c}{2min(a,\ d)+b+c}$	1954	$<-1,+1>$
29	Scott[8]	$\frac{4ad-(b+c)^2}{(2a+b+c)(2d+b+c)}$	1955	$<-1,+1>$
30	Ochiai[7,9]	$\frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	1957	$<0,+1>$
31	Sokal/Michener[5-7,9]	$\frac{a+d}{n}$	1958	$<0,+1>$
32	Sorgenfrei[5,7,9]	$\frac{a^2}{(a+b)(a+c)}$	1959	$<0,+1>$
33	Cohen[8]	$\frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}$	1960	$<-1,+1>$
34	Rogers/Tanimoto I[6,7]	$\frac{a+d}{a+2b+2c+d}$	1960	$<0,+1>$
35	Rogers/Tanimoto II[5,9]	$\frac{a+d}{a+2(b+c+d)}$	1960	$<0,+1>$
36	Tarwid[5,7,9]	$\frac{[an-(a+b)(a+c)]}{[an+(a+b)(a+c)]}$	1960	$<-1,+1>$
37	Hamann[5-7,9]	$\frac{a+d-b-c}{n}$	1961	$<-1,+1>$
38	Stiles[6,7,9]	$log_{10}\frac{n\left(\lvert ad-bc\rvert-\frac{n}{2}\right)^2}{(a+b)(a+c)(b+d)(c+d)}$	1961	$(-\infty,+\infty)$
39	Mountford[5,7,9]	$\frac{2a}{ab+ac+2bc}$	1962	$<0,+2>$
40	Preston[11]	$\sqrt{\frac{a+b}{a+b+c}}+\sqrt{\frac{a+c}{a+b+c}}$	1962	$<+1,+2>$
41	Sokal/Sneath I[5-7,9]	$\frac{a}{a+2b+2c}$	1963	$<0,+1>$
42	Sokal/Sneath II[5-7,9]	$\frac{2a+2d}{2a+b+c+2d}$	1963	$<0,+1>$
43	Sokal/Sneath III[5,7,9]	$\frac{1}{4}\left[\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{b+d}+\frac{d}{c+d}\right]$	1963	$<0,+1>$
44	Sokal/Sneath IV[5,7,9]	$\frac{a}{\sqrt{(a+b)(a+c)}}\frac{d}{\sqrt{(b+d)(c+d)}}$	1963	$<0,+\infty)$
45	Sokal/Sneath V[5-7]	$\frac{a+d}{b+c}$	1963	$<0,+\infty)$
46	Fager/McGowan[5,7,9]	$\frac{a}{\sqrt{[(a+b)(a+c)]}}-\frac{1}{2}[max\{(a+b),(a+c)\}]$	1963	$(-\infty,+0{,}5)$
47	McConnaughey[5-7,9]	$\frac{a^2-bc}{(a+b)(a+c)}$	1964	$<-1,+1>$
48	Dennis[6,7,9]	$\frac{ad-bc}{\sqrt{n(a+b)(a+c)}}$	1965	$(-\infty,+\infty)$
49	Fossum[6,7,9]	$\frac{n(a-0.5)^2}{(a+b)(a+c)}$	1966	$<0,+\infty)$
50	Rogot/Goldberg[8]	$\frac{a}{2a+b+c}+\frac{d}{2d+b+c}$	1966	$<0,+1>$
51	Gilbert/Wells[5,7,9]	$\log a-\log n-log\left[\frac{(a+b)}{n}\right]-log\left[\frac{(a+c)}{n}\right]$	1966	$<-3,+2>$
52	Johnson[5,7,9]	$\frac{a}{a+b}+\frac{a}{a+c}$	1967	$<0,+2>$

TABLO 2: Benzerlik katsayıları *(devamı)*.

No	Adı	Formül	Kronoloji	Değer aralığı
53	Goodall[5]	$(50\pi)^{-1} arcsin\sqrt{\frac{a+d}{n}}$	1967	$< 0, +0.01 >$
54	Hawkins/Dotson[8]	$\frac{1}{2}\left(\frac{a}{a+b+c}+\frac{d}{b+c+d}\right)$	1968	$< 0, +1 >$
55	Maxwell/Pilliner[8]	$\frac{2(ad-bc)}{(a+b)(c+d)+(a+c)(b+d)}$	1968	$< -1, +1 >$
56	Hurlbert[5]	$\sqrt{\frac{\chi^2}{\chi^2_{max}}}$ $\chi^2_{max}=\frac{n(a+b)(b+d)}{(a+c)(c+d)}\,;ad \geq bc$ $\chi^2_{max}=\frac{n(a+b)(a+c)}{(b+d)(c+d)}\,;ad < bc \; and \; a \leq d$ $\chi^2_{max}=\frac{n(b+d)(c+d)}{(a+b)(a+c)}\,;ad < bc \; and \; a > d$	1969	$< 0, +1 >$
57	van der Maarel[8]	$\frac{2a-b-c}{2a+b+c}$	1969	$< -1, +1 >$
58	Gower[7,9]	$\frac{a+d}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	1971	$< 0, +1{,}11)$
59	Anderberg[7,9]	$\frac{\sigma-\sigma'}{2n}$ $\sigma = max(a,b)+max(c,d)+max(a,c)+max(b,d)$ $\sigma' = max(a+c,b+d)+max(a+b,c+d)$	1973	$< 0, +0{,}5 >$
60	Baroni Urbani/Buser I[5,7,9]	$\frac{\sqrt{ad}+a-b-c}{\sqrt{ad}+a+b+c}$	1976	$< -1, +1 >$
61	Baroni Urbani/Buser II[5-7,9]	$\frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c}$	1976	$< 0, +1 >$
62	Austin-Colwell[8]	$\frac{2}{\pi} arcsin\sqrt{\left(\frac{a+d}{n}\right)}$	1977	$< 0, +1{,}005 >$
63	Harris-Lahey[8]	$\frac{a(2d+b+c)}{2(a+b+c)}+\frac{d(2a+b+c)}{2(b+c+d)}$	1978	$< 0, +\infty)$
64	Faith[7,9]	$\frac{a+0.5d}{n}$	1983	$< 0, +1 >$
65	Digby[12]	$\frac{(ad)^{3/4}-(bc)^{3/4}}{(ad)^{3/4}+(bc)^{3/4}}$	1983	$< -1, +1 >$
66	Choi et al.[7,8]	$\frac{ad-bc}{n^2}$	2012	$(-1, +1)$
67	Consonni/Todeschini I[13]	$\frac{ln(1+a+d)}{ln(1+n)}$	2012	$< 0, +1 >$
68	Consonni/Todeschini II[13]	$\frac{ln(1+n)-ln(1+b+c)}{ln(1+n)}$	2012	$< 0, +1 >$
69	Consonni/Todeschini III[13]	$\frac{ln(1+a)}{ln(1+n)}$	2012	$< 0, +1 >$
70	Consonni/Todeschini IV[13]	$\frac{ln(1+a)}{ln(1+a+b+c)}$	2012	$< 0, +1 >$
71	Consonni/Todeschini V[13]	$\frac{ln(1+ad)-ln(1+bc)}{ln\left((1+n^2/4)\right)}$	2012	$< -1, +1 >$
72	Tarantula[7,9]	$\frac{a(c+d)}{c(a+b)}$	?	$< 0, +\infty)$

TABLO 3: $n, a, b, c\ ve\ d$'nin farklı değerleri için elde edilen sonuçlar.

	$a = n$	$b = n$	$c = n$	$d = n$	$a + d = n$	$b + c = n$	$a = b = c = d$	$a + d$
Peirce I	NC	NC	NC	NC	1	-1	0	C
Peirce II	NC	NC	NC	NC	1	-1	0	C
Doolittle	NC	NC	NC	NC	1	1	0	C-
Yule I	NC	NC	NC	NC	1	-1	0	C
Jaccard I	1	0	0	NC	1	0	0,333	C
Pearson I	NC	NC	NC	NC	n	n	0	C-
Pearson II	NC	NC	NC	NC	0,707	0,707	0	C-
Forbes I	1	NC	NC	NC	2	0	1	C
Jaccard II	1	0	0	NC	1	0	0,6	C
Yule II	NC	NC	NC	NC	1	-1	0	C
Yule III	NC	NC	NC	NC	1	-1	0	C
Czekanowski	1	0	0	NC	1	0	0,5	C
Pearson-Heron	NC	NC	NC	NC	1	-0,598	-0,448	C
Michael	0	0	0	0	1	-1	0	C
Forbes II	NC	NC	NC	NC	1	-1	0	C-
Kulczynski I	1	NC	NC	NC	1	0	0,5	C
Kulczynski II	NC	0	0	NC	NC	0	0,5	C
Braun-Blanquet	1	0	0	NC	1	0	0,5	C
Driver-Kroeber	1	NC	NC	NC	1	0	0,5	C
Eyraud	NC	NC	NC	NC	-0,032	-0,04	C+	C
Russell-Rao	1	0	0	0	0,5	0	0,25	C
Simpson	1	NC	NC	NC	1	0	0,5	C
Dice I	1	0	NC	NC	1	0	0,5	C
Dice II	1	NC	0	NC	1	0	0,5	C
Cole I	NC	NC	0	NC	1	-1	0	C
Cole II	NC	0	NC	NC	1	-1	0	C
Cole III	NC	NC	NC	NC	1	-1	0	C
Goodman-Kruskal	NC	-1	-1	NC	1	-1	0	C
Scott	NC	-1	-1	NC	1	-1	0	C
Ochiai	NC	NC	NC	NC	1	0	0,25	C
Sokal-Michener	1	0	0	1	1	0	0,5	C+
Sorgenfrei	1	NC	NC	NC	1	0	0,25	C
Cohen	NC	0	0	NC	1	-1	0	C
Rogers-Tanimoto I	1	0	0	1	1	0	0,333	C+
Rogers-Tanimoto II	1	0	0	0,5	0,667	0	0,285	C
Tarwid	0	NC	NC	NC	0,333	-1	0	C
Hamann	1	-1	-1	1	1	-1	0	C+
Stiles	NC	NC	NC	NC	0,806	0,806	C	C-
Mountford	NC	NC	NC	NC	NC	0	C	C+
Preston	2	1	1	NC	2	1,414	1,633	C
Sokal-Sneath I	1	0	0	NC	1	0	0,2	C
Sokal-Sneath II	1	0	0	1	1	0	0,667	C+
Sokal-Sneath III	NC	NC	NC	NC	1	0	0,5	C

TABLO 3: $n, a, b, c\ ve\ d$'nin farklı değerleri için elde edilen sonuçlar *(devamı).*

	$a = n$	$b = n$	$c = n$	$d = n$	$a + d = n$	$b + c = n$	$a = b = c = d$	$a + d$
Sokal-Sneath IV	NC	NC	NC	NC	5	0	C	C
Sokal-Sneath V	NC	0	0	NC	NC	0	1	C+
Fager-McGowan	-4	NC	NC	NC	-1,5	-2,5	C	C
McConnaughey	1	NC	NC	NC	1	-1	0	C
Rogot-Goldberg	NC	0	0	NC	1	0	0,5	C
Gilbert-Wells	0	NC	NC	NC	0,301	NC	0	C
Johnson	2	NC	NC	NC	2	0	1	C
Goodall	0,01	0	0	0,01	0,01	0	0,005	C+
Hawkins-Dotson	NC	0	0	NC	1	0	0,333	C
Maxwell-Pilliner	NC	NC	NC	NC	1	-1	0	C
Hurlbert	NC	NC	NC	NC	1	1	0	C-
van der Maarel	1	-1	-1	NC	1	-1	0	C
Gower	NC	NC	NC	NC	0,4	0	C	C
Anderberg	0	0	0	0	0,5	0,5	0	C-
Baroni Urbani-Buser I	1	-1	-1	NC	1	-1	0	C
Baroni Urbani-Buser II	1	0	0	NC	1	0	0,5	C
Austin-Colwell	1	0	0	1	1	0	0,500	C+
Harris-Lahey	NC	0	0	NC	n	0	C	C
Faith	1	0	0	0,5	0,75	0	0,375	C
Digby	NC	NC	NC	NC	1	-1	0	C
Dennis et al.	0	NC	NC	NC	1,581	-1,581	0	C
Fossum et al.	9,025	NC	NC	NC	8,1	0,1	C	C
Choi et al.	0	0	0	0	0,25	-0,25	0	C
Consonni-Todeschini I	1	0	0	1	1	0	C	C+
Consonni-Todeschini II	1	0	0	1	1	0	C	C+
Consonni-Todeschini III	1	0	0	0	0,747	0	C	C
Consonni-Todeschini IV	1	0	0	NC	1	0	C	C
Consonni-Todeschini V	0	0	0	0	1	-1	0	C
Tarantula	NC	NC	NC	NC	NC	0	1	C

NC: Hesaplanamaz; C: $n, a, b, c\ ve\ d$ değerlerine bağlı olarak farklı değerler alır; C+: $(a + d)$ değeri arttıkça değeri büyür; C-: $(a + d)$ değeri belirli bir değere ulaşıncaya kadar değeri küçülür, sonra değeri büyür.

TABLO 4: Hiyerarşik Kümeleme Analizi sonuçları.

n	Küme no	Yöntemler
10	1	Anderberg, Austin-Colwell, Baroni Urbani-Buser I, Baroni Urbani-Buser II, Braun-Blanquet, Consonni-Todeschini I, Consonni-Todeschini II, Consonni-Todeschini III, Consonni-Todeschini IV, Consonni-Todeschini V, Choi, Cohen, Cole I, Cole II, Cole III, Czekanowski, Dennis, Dice I, Dice II, Digby, Doolittle, Driver-Kroeber, Eyraud, Faith, Forbes II, Gilbert-Wells, Goodall, Goodman-Kruskal, Gower, Hamann, Hawkins-Dotson, Hurlbert, Jaccard I, Jaccard II, Kulczynski I, Maxwell-Pilliner, McConnaughey, Michael, Mountford, Ochiai, Pearson II, Pearson-Heron, Peirce I, Peirce II, Rogers-Tanimoto I, Rogers-Tanimoto II, Rogot-Goldberg, Russell-Rao, Scott, Simpson, Sokal-Michener, Sokal-Sneath I, Sokal-Sneath II, Sokal-Sneath III, Sorgenfrei, Tarwid, van der Maarel, Yule I, Yule II, Yule III
	2	Forbes I, Johnson, Preston
	3	Stiles
	4	Kulczynski II
	5	Sokal-Sneath IV
	6	Sokal-Sneath V
	7	Tarantula
	8	Harris-Lahey
	9	Fossum
	10	Pearson I
	11	Fager-McGowan
15	1	Anderberg, Austin-Colwell, Baroni Urbani-Buser I, Baroni Urbani-Buser II, Braun-Blanquet, Consonni-Todeschini I, Consonni-Todeschini II, Consonni-Todeschini III, Consonni-Todeschini IV, Consonni-Todeschini V, Choi, Cohen, Cole I, Cole II, Cole III, Czekanowski, Dennis, Dice I, Dice II, Digby, Doolittle, Driver-Kroeber, Eyraud, Faith, Forbes II, Gilbert-Wells, Goodall, Goodman-Kruskal, Gower, Hamann, Hawkins-Dotson, Hurlbert, Jaccard I, Jaccard II, Kulczynski I, Maxwell-Pilliner, McConnaughey, Michael, Mountford, Ochiai, Pearson II, Pearson-Heron, Peirce I, Peirce II, Rogers-Tanimoto I, Rogers-Tanimoto II, Rogot-Goldberg, Russell-Rao, Scott, Simpson, Sokal-Michener, Sokal-Sneath I, Sokal-Sneath II, Sokal-Sneath III, Sorgenfrei, Tarwid, van der Maarel, Yule I, Yule II, Yule III
	2	Forbes I, Johnson, Preston
	3	Kulczynski II, Sokal-Sneath IV
	4	Stiles
	5	Sokal-Sneath V
	6	Tarantula
	7	Harris-Lahey
	8	Fossum
	9	Pearson I
	10	Fager-McGowan
20	1	Anderberg, Austin-Colwell, Baroni Urbani-Buser I, Baroni Urbani-Buser II, Braun-Blanquet, Consonni-Todeschini I, Consonni-Todeschini II, Consonni-Todeschini III, Consonni-Todeschini IV, Consonni-Todeschini V, Choi, Cohen, Cole I, Cole II, Cole III, Czekanowski, Dennis, Dice I, Dice II, Digby, Doolittle, Driver-Kroeber, Eyraud, Faith, Forbes I, Forbes II, Gilbert-Wells, Goodall, Goodman-Kruskal, Gower, Hamann, Hawkins-Dotson, Hurlbert, Jaccard I, Jaccard II, Johnson, Kulczynski I, Maxwell-Pilliner, McConnaughey, Michael, Mountford, Ochiai, Pearson II, Pearson-Heron, Peirce I, Peirce II, Preston, Rogers-Tanimoto I, Rogers-Tanimoto II, Rogot-Goldberg, Russell-Rao, Scott, Simpson, Sokal-Michener, Sokal-Sneath I, Sokal-Sneath II, Sokal-Sneath III, Sorgenfrei, Stiles, Tarwid, van der Maarel, Yule I, Yule II, Yule III
	2	Kulczynski II, Sokal-Sneath IV, Sokal-Sneath V, Tarantula
	3	Harris-Lahey
	4	Fossum
	5	Pearson I
	6	Fager-McGowan

TABLO 4: Hiyerarşik Kümeleme Analizi sonuçları *(devamı)*.

n	Küme no	Yöntemler
25, 30, 40, 45, 55	1	Anderberg, Austin-Colwell, Baroni Urbani-Buser I, Baroni Urbani-Buser II, Braun-Blanquet, Consonni-Todeschini I, Consonni-Todeschini II, Consonni-Todeschini III, Consonni-Todeschini IV, Consonni-Todeschini V, Choi, Cohen, Cole I, Cole II, Cole III, Czekanowski, Dennis, Dice I, Dice II, Digby, Doolittle, Driver-Kroeber, Eyraud, Faith, Forbes I, Forbes II, Gilbert-Wells, Goodall, Goodman-Kruskal, Gower, Hamann, Hawkins-Dotson, Hurlbert, Jaccard I, Jaccard II, Johnson, Kulczynski I, Kulczynski II, Maxwell-Pilliner, McConnaughey, Michael, Mountford, Ochiai, Pearson II, Pearson-Heron, Peirce I, Peirce II, Preston, Rogers-Tanimoto I, Rogers-Tanimoto II, Rogot-Goldberg, Russell-Rao, Scott, Simpson, Sokal-Michener, Sokal-Sneath I, Sokal-Sneath II, Sokal-Sneath III, Sorgenfrei, Stiles, Tarwid, van der Maarel, Yule I, Yule II, Yule III
	2	Sokal-Sneath V, Tarantula
	3	Sokal-Sneath IV
	4	Harris-Lahey
	5	Fossum
	6	Pearson I
	7	Fager-McGowan
35, 50, □ ≥ *60*	1	Anderberg, Austin-Colwell, Baroni Urbani-Buser I, Baroni Urbani-Buser II, Braun-Blanquet, Consonni-Todeschini I, Consonni-Todeschini II, Consonni-Todeschini III, Consonni-Todeschini IV, Consonni-Todeschini V, Choi, Cohen, Cole I, Cole II, Cole III, Czekanowski, Dennis, Dice I, Dice II, Digby, Doolittle, Driver-Kroeber, Eyraud, Faith, Forbes I, Forbes II, Gilbert-Wells, Goodall, Goodman-Kruskal, Gower, Hamann, Hawkins-Dotson, Hurlbert, Jaccard I, Jaccard II, Johnson, Kulczynski I, Kulczynski II, Maxwell-Pilliner, McConnaughey, Michael, Mountford, Ochiai, Pearson II, Pearson-Heron, Peirce I, Peirce II, Preston, Rogers-Tanimoto I, Rogers-Tanimoto II, Rogot-Goldberg, Russell-Rao, Scott, Simpson, Sokal-Michener, Sokal-Sneath I, Sokal-Sneath II, Sokal-Sneath III, Sokal-Sneath V, Sorgenfrei, Stiles, Tarantula, Tarwid, van der Maarel, Yule I, Yule II, Yule III
	2	Sokal-Sneath IV
	3	Harris-Lahey
	4	Fossum
	5	Pearson I
	6	Fager-McGowan

TARTIŞMA

Literatürde az sayıda da olsa benzerlik katsayılarının karşılaştırıldığı çalışmalar bulunmaktadır. Ancak bu çalışmalarda genellikle ölçütlerin sınıflandırılması üzerinde durulmuş ve hangi benzerlik ölçütlerinin bir araya geldiği belirlenmeye çalışılmıştır. Bu çalışmalarda ölçütler genel olarak, negatif benzerliği doğrudan veya farklı şekillerde ağırlıklandırılmış hâlini dikkate alan ve negatif benzerliği dikkate almayan ölçütler olarak 2 gruba ayrılmaktadır. Ölçütlerin büyük bir kısmı hesaplamalarda $a, b, c\ ve\ d$ değerlerini kullanmasına rağmen hesaplamalarda d'yi tamamen hariç tutan veya d'yi içeren ancak a'ya göre ölçüte katkısını azaltan ölçütler de bulunmaktadır. d'nin hesaplamalarda dikkate alınması, özellikle ekolojik ve biyolojik verilerde bilgilendirici olma özelliğinden dolayı özellikle ön plana çıkmaktadır. Choi ve ark. tarafından ise d'nin, nitelikler arasında hiçbir benzerliği yansıtmayabileceği öne sürülmüştür.[7] Wijaya ve ark., tarafından benzerlik ölçütlerinin tamamının benzerliği hesaplamak için önemli olduğu, d'yi içeren ikili benzerlik ölçütlerinin, d'yi hariç tutan ölçütlere kıyasla benzerliği belirlemede daha iyi oldukları ifade edilmiştir.[9] Brusco ve ark. tarafından yapılan çalışmada ise benzerlik ölçütleri 2 farklı kategoriye ayrılmıştır. İlk grupta yer alan ölçütler tipik olarak 0 ile 1 arasında değişir ve paylarında genellikle a veya $a + d$ bulunur. Bu grupta yer alan ölçütler $a, b, c\ ve\ d$ değerlerinden önemli derecede etkilenir. Buna karşılık, 2. grupta yer alan ölçütler tipik olarak -1 ile +1 arasında değişir ve paylarında genellikle $ad - bc$ bulunur. İkinci grupta yer alan ölçütlerin birinci grupta yer alan ölçütlere nazaran $a, b, c\ ve\ d$ değerlerinden daha az etkilenmelerini sağlayan doğal bir mer-

kezleme etkisine sahip oldukları ifade edilmektedir.[10] Çalışmadan elde edilen sonuçlar ile literatürde ifade edilen bu durumlar benzerlik göstermektedir.

SONUÇ

İkili veriler için benzerlik katsayılarının $0 \leq Benzerlik\ Katsayısı \leq 1$ aralığında değer alması beklenmesine rağmen tüm katsayılar için bu aralık geçerli değildir. Dikkate alınan 72 farklı katsayı içerisinden yalnızca 29 tanesi bu aralıkta değer almaktadır. Hiyerarşik Kümeleme Analizi'ne göre benzerlik katsayılarının çoğu birbirine benzemektedir. Fager-McGowan, Fossum, Harris-Lahey, Pearson I ve Sokal-Sneath IV katsayıları tüm n değerleri için diğer katsayılardan farklılık göstermektedir. Eyraud, Fager-McGowan, Fossum, Gower, Harris-Lahey, Pearson I ve Sokal-Sneath IV benzerlik katsayıları n sayısından etkilenmekte diğer katsayılar etkilenmemektedir. Dolayısıyla benzerlik katsayılarının önemli bir kısmının örnek büyüklüğünden bağımsız oldukları belirlenmiştir. Genel olarak hemen hemen tüm katsayılara ait değerler, örnekler daha benzer hâle geldikçe sabit bir minimumdan sabit bir maksimuma doğru artmaktadır. Hurlbert, Doolitle, Pearson I, Pearson II, Forbes II, Stiles ve Anderberg katsayıları, $a + d$ değeri belirli bir değere ulaşana kadar azalmakta, değere ulaştıktan sonra ise artmaktadır. $a + d$ değeri arttıkça Rogers-Tanimoto I, Mountford, Sokal-Sneath II, Sokal-Sneath V, Goodall, Austin-Colwell, Consonni-Todeschini I ve Consonni-Todeschini II katsayılarının değerleri doğrusal olmayan artış göstermekte, Hamann ve Sokal-Michener katsayıları ise tüm n değerleri için benzerlik ile doğrusal olarak artmaktadır. Değer aralığının $0 - 1$ olması ve benzerlik artışı ile paralellik göstermesinden dolayı Sokal-Michener katsayısı tüm katsayılar içerisinde öne çıkmaktadır.

Finansal Kaynak

Bu çalışma sırasında, yapılan araştırma konusu ile ilgili doğrudan bağlantısı bulunan herhangi bir ilaç firmasından, tıbbi alet, gereç ve malzeme sağlayan ve/veya üreten bir firma veya herhangi bir ticari firmadan, çalışmanın değerlendirme sürecinde, çalışma ile ilgili verilecek kararı olumsuz etkileyebilecek maddi ve/veya manevi herhangi bir destek alınmamıştır.

Çıkar Çatışması

Bu çalışma ile ilgili olarak yazarların ve/veya aile bireylerinin çıkar çatışması potansiyeli olabilecek bilimsel ve tıbbi komite üyeliği veya üyeleri ile ilişkisi, danışmanlık, bilirkişilik, herhangi bir firmada çalışma durumu, hissedarlık ve benzer durumları yoktur.

Yazar Katkıları

Fikir/Kavram: *İsmet Doğan, Nurhan Doğan;* ***Tasarım:*** *İsmet Doğan, Nurhan Doğan;* ***Denetleme/Danışmanlık:*** *İsmet Doğan, Nurhan Doğan, Taylan Doğan;* ***Veri Toplama ve/veya İşleme:*** *İsmet Doğan, Nurhan Doğan, Taylan Doğan;* ***Analiz ve/veya Yorum:*** *İsmet Doğan, Nurhan Doğan, Taylan Doğan;* ***Kaynak Taraması:*** *İsmet Doğan, Nurhan Doğan;* ***Makalenin Yazımı:*** *İsmet Doğan, Nurhan Doğan, Taylan Doğan;* ***Eleştirel İnceleme:*** *İsmet Doğan, Nurhan Doğan, Taylan Doğan;* ***Kaynaklar ve Fon Sağlama:*** *İsmet Doğan, Nurhan Doğan;* ***Malzemeler:*** *İsmet Doğan, Nurhan Doğan.*

KAYNAKLAR

1. Wolda H. Similarity indices, sample size and diversity. Oecologia. 1981;50(3):296-302. [Crossref] [PubMed]
2. Wong KS, Kim MH. Privacy-preserving similarity coefficients for binary data. Comput Math Appl. 2013;65(9):1280-90. [Crossref]
3. Willett P. Similarity-based approaches to virtual screening. Biochem Soc Trans. 2003;31(Pt 3):603-6. [Crossref] [PubMed]
4. Cheetham AH, Hazel JE. Binary (presence-absence) similarity coefficients. J Paleontol. 1969;43(5):1130-6. [Link]
5. Hubalek Z. Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. Biol Rev. 1982;57:669-89. [Crossref]
6. Holliday JD, Hu CY, Willett P. Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. Comb Chem High Throughput Screen. 2002;5(2):155-66. [Crossref] [PubMed]
7. Choi SS, Cha SH, Tappert CC. A survey of binary similarity and distance measures. J Syst Cybern Inf. 2010;8(1):43-8. [Link]
8. Todeschini R, Consonni V, Xiang H, Holliday J, Buscema M, Willett P. Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets. J Chem Inf Model. 2012;52(11):2884-901. [Crossref] [PubMed]
9. Wijaya SH, Afendi FM, Batubara I, Darusman LK, Altaf-Ul-Amin M, Kanaya S. Finding an appropriate equation to measure similarity between binary vectors: case studies on Indonesian and Japanese herbal medicines. BMC Bioinformatics. 2016;17(1):520. [Crossref] [PubMed] [PMC]
10. Brusco M, Cradit JD, Steinley D. A comparison of 71 binary similarity coefficients: The effect of base rates. PLoS One. 2021;16(4):e0247751. [Crossref] [PubMed] [PMC]
11. Peters JA. A computer program for calculating degree of biogeographical resemblance between areas. Syst Zool. 1968;17(1):64-9. [Crossref]
12. Warrens MJ. On association coefficients for 2x2 tables and properties that do not depend on the marginal distributions. Psychometrika. 2008;73(4):777-89. [Crossref] [PubMed] [PMC]
13. Consonni V, Todeschini R. New similarity coefficients for binary data. MATCH Commun Math Comput Chem. 2012;68:581-92. [Link]