Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2022;14(2):63-9

ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH DOI: 10.5336/biostatic.2022-88302

İki Sonuçlu Nitel Veriler İçin Farklılık/Uzaklık Katsayılarının Değerlendirilmesi: Bir Benzetim Çalışması

Evaluation of Dissimilarity/Distance Coefficients of Binary Data: A Simulation Study

İsmet DOĞAN[a], Nurhan DOĞAN[a]

[a]Afyonkarahisar Sağlık Bilimleri Üniversitesi Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim ABD, Afyonkarahisar, Türkiye

ÖZET Amaç: Bu çalışmanın amacı, 2 sonuçlu veriler ile ilgili türetilmiş veri setleri kullanarak farklı n, a, b, c ve d değerleri için belirlenen 23 farklı uzaklık katsayısını tanıtmak, özelliklerini ortaya koyarak değerlendirmektir. **Gereç ve Yöntemler:** Bu çalışmada, 2 sonuçlu veriler için ileri sürülen uzaklık 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 286, $n = 15$ için 815 ve $n \geq 20$ için biner farklı veri seti çalışmada kullanılmıştır. **Bulgular:** İki sonuçlu veriler için tüm farklılık/uzaklık katsayılarının değer aralığının 0 (benzerlik yok) ile 1 (tam benzerlik) olması beklenmesine rağmen tüm katsayılar için bu aralık geçerli değildir. Dikkate alınan 23 farklı katsayı içerisinden 12 tanesi bu aralıkta değer almaktadır. Hiyerarşik kümeleme analizine göre farklılık/uzaklık 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 Sokal-Michener, Hamming ve varyans katsayıları, tüm n değerleri için farklılık/uzaklık ile doğrusal olarak sorunsuz bir şekilde artmaktadır. Değer aralığının 0-1 olması ve farklılık/uzaklık artışı ile paralellik göstermesinden dolayı Sokal-Michener tarafından önerilen katsayı tüm katsayılar içerisinde öne çıkmaktadır. Cosine, Hamming, Euclid I ve Euclid II katsayıları n sayısından etkilenmekte diğer katsayılar etkilenmemektedir. Dolayısıyla farklılık/uzaklık katsayılarının önemli bir kısmının örnek büyüklüğünden bağımsız oldukları belirlenmiştir.

Anahtar kelimeler: Farklılık/uzaklık katsayısı; hiyerarşik kümeleme; iki sonuçlu veri

ABSTRACT Objective: The aim of this study is to introduce 23 different binary dissimilarity/distance 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 dissimilarity/distance 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. 286 for $n = 10$, 815 for $n = 15$ and 1000 different data sets for $n \geq 20$ were used in the study. **Results:** Although the value range of all dissimilarity/distance coefficients for binary data is expected to be 0 to 1, this range is not valid for all coefficients. Out of 23 different coefficients, 12 take values in this range. According to the hierarchical clustering analysis, most of the dissimilarity/distance 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 dissimilar. However, the Sokal-Michener, Hamming and variance coefficients increase linearly with dissimilarity/distance smoothly for all n values. The coefficient suggested by Sokal-Michener stands out among all the coefficients because the value range is 0-1 and is parallel to the dissimilarity/distance increase. Cosine, Hamming, Euclid I and Euclid II coefficients are affected by the number of n and other coefficients are not. Therefore, it has been determined that a significant part of the dissimilarity/distance coefficients are independent of the sample size.

Keywords: Dissimilarity/distance coefficient; hierarchical clustering; binary data

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

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 20 Jan 2022 ***Received in revised form:*** 18 Apr 2022 ***Accepted:*** 09 May 2022 ***Available online:*** 08 Jun 2022

Bilimsel ve matematiksel açıdan uzaklık, 2 nesnenin birbirinden ne kadar uzak olduğunun nicel bir derecesi olarak tanımlanmakta, metrik özellikleri (negatif olmama, güçlü refleksivite, simetri, üçgen eşitsizliği) karşılayan uzaklık ölçüleri basitçe metrik olarak, diğer uzaklık ölçüleri ise fikir ayrılığı/uyuşmazlık ölçüleri olarak isimlendirilmektedir. Öklid uzaklığı yaygın olarak kullanılmasına rağmen karşılaştırılacak her tür veri veya model için uygun değildir. XX. yüzyılda, çeşitli uygulamalar için yeni uzaklık ölçütlerinden yararlanmaya yönelik muazzam çabalar söz konusu olmuştur. Antropoloji, biyoloji, kimya, bilgisayar bilimi, ekoloji, bilgi teorisi, jeoloji, matematik, fizik, psikoloji, istatistik gibi birçok farklı alanda karşılaşılan önemli sayıda uzaklık ölçüleri bulunmaktadır.[1] Veri madenciliği ve makine öğreniminde veri nesneleri arasındaki farklılığı veya benzerliği ölçmek, özellikle uzaklık tabanlı kümeleme ve uzaklık tabanlı sınıflandırma gibi uzaklık tabanlı tekniklerin birincil görevlerinden biridir. Ancak kategorik verilerdeki benzerlik ya da farklılığın ölçülmesi zorlu bir problemdir çünkü kategorik veriler herhangi bir yapıya sahip değildir ve bu nedenle yalnızca özdeş bir karşılaştırma işlemi uygulanabilir.[2] Benzerlik kavramının tamamlayıcısı "farklılık" kavramıdır. Ancak farklılık kavramı ile birlikte uzaklık terimi de sıklıkla kullanılmaktadır. Benzerlik, nitel/nicel özellikler için eşleşme/örtüşme açısından, farklılık ise uyumsuzluk/fark açısından tanımlanır. Araştırmacılar tarafından sıklıkla $Farklılık = 1 - Benzerlik$ olarak düşünülse de bu durum sadece değer aralığı 0-1 olan katsayılar için geçerlidir. Ancak benzerlik katsayılarının tamamı bu aralıkta değer almamaktadır. Dolayısıyla $Farklılık = 1 - Benzerlik$ yaklaşımından bağımsız olarak literatüre kazandırılmış çok sayıda farklılık/uzaklık katsayısı bulunmaktadır. Bilim insanları yaptıkları çalışmalarda, benzerlik ve farklılık/uzaklık kavramları hakkında titiz davranmaktadır. Uzaklık katsayıları, çok boyutlu geometrik uzaydaki uzaklıklara benzerdir, ancak bu tür uzaklıklara tam olarak eş değer olmaları gerekmez. Bir fonksiyon ancak belirli gereksinimleri veya aksiyomları karşılıyorsa benzerlik veya farklılık/uzaklık olarak kabul edilebilir. Benzerlik gereksinimlerine benzer şekilde, farklılık/uzaklık kavramı için de aksiyomlar vardır. Farklılıklar/uzaklıklar, çeşitli çok değişkenli veri analizi yöntemleriyle değerlendirilen işlevlerdir. İyi bilinen örnekler; çok boyutlu ölçekleme ve kümeleme analizidir.[3] Klasik bir 2×2 tablo örneği Tablo 1'de verilmiştir.

TABLO 1: Farklılık/uzaklık 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

Farklılık/uzaklık 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. İkili veriler için bugüne kadar literatürde yer alan, her biri kendi matematiksel özelliklerine sahip olan ve farklı bilimsel alanlarda kullanılan 23 farklı farklılık/uzaklık katsayısı belirlenmiştir. Bazı katsayıların literatürde farklı isimlerle yer alması nedeniyle seçim sürecinde, aynı olan katsayıların (örneğin Jaccard-Needham katsayısı, Soergel katsayısı, Bray-Curtis katsayısı, Lance-Williams katsayısı, Sorensen katsayısı ya da Watson katsayısı olarak bilinir) hariç tutulmasına gayret edilmiştir. İki sonuçlu nitel veriler ile nicel 2 değişkene ait veriler için önerilen uzaklık katsayıları karşılaştırıldığında;

- İki sonuçlu nitel veriler için önerilen Hamming uzaklığının, nicel 2 değişkene ait veriler için önerilen Manhattan uzaklığı ile
- İki sonuçlu nitel veriler için önerilen Hamming uzaklığının karekökünün, nicel 2 değişkene ait veriler için önerilen Öklid uzaklığının karesi ile

- İki sonuçlu nitel veriler için önerilen Tanimoto uzaklığının, nicel 2 değişkene ait veriler için önerilen ortalama Manhattan uzaklığı ile

- İki sonuçlu nitel veriler için önerilen Tanimoto uzaklığının, karesinin nicel 2 değişkene ait veriler için önerilen ortalama Öklid uzaklığı ile çakıştığını görmek kolaydır. Ayrıca Bray-Curtis uzaklık katsayısı, Gleason benzerlik katsayısının, Jaccard-Needham uzaklık katsayısı ise Jaccard-Tanimoto benzerlik katsayısının tümleyenidir.[4] Çalışmada, 2'li veriler için belirlenen 23 farklılık/uzaklık katsayısı tanıtılmış ve birbirleriyle karşılaştırılmıştır. Bu çalışmanın amacı, türetilmiş veri setleri kullanarak farklı değerleri için belirlenen 23 farklı 2 sonuçlu farklılık/uzaklık katsayısını tanıtmak, özelliklerini ortaya koyarak değerlendirmektir. Makalede, Helsinki Deklarasyonu Prensipleri dikkate alınmıştır.

GEREÇ VE YÖNTEMLER

Bu çalışmada, özellikle 2 sonuçlu veriler için kullanılması önerilen farklılık/uzaklık katsayıları ele alınmıştır, çünkü 2 sonuçlu veriler, çeşitli verileri temsil etmek için yaygın olarak kullanılmaktadır. İki sonuçlu veriler için geliştirilmiş farklılık/uzaklık katsayıları evet/hayır, doğru/yanlış veya var/yok biçiminde veriler mevcut olduğunda kullanılır ve bu nedenle nominal ölçek için uygundur. 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. Veri türetimine ilişkin tarafımızdan yazılan programda yer alan simülasyon kurgusunun detayları aşağıdaki gibidir:

Adım 1. Veri türetimi için gerekli değerler (toplam kaç adet veri türetilmek istendiği, kategorilere atanacak değerlerin toplamı, kategori sayısı) programa girilir.

Adım 2. Python random kütüphanesindeki choice fonksiyonu kullanılarak rastgele bir şekilde kategori seçimi yapılır.

Adım 3. Choice fonksiyonu tekrar kullanılarak seçilen kategori için 0 ile istenen toplam arasında rastgele bir tam sayı seçimi yapılır. Örneğin istenen toplam 30 ise [0, 30] aralığında rastgele bir tam sayı seçilir.

Adım 4. Tüm kategoriler için 2 ve 3. adımlar tekrarlanır (k'ıncı kategori için bir tam sayı seçildikten sonra seçilen sayıların toplamına bakılır. Eğer bu toplam, istenilen toplamdan fazla ise bu geçersiz bir veri olacağından işleme baştan başlanır. Son kategoriye atanacak tam sayı ise istenilen toplamdan mevcut toplamın çıkarılmasıyla bulunur.

Adım 5. Türetilen verinin daha önce türetilip türetilmediğine bakılır. Aynı veri daha önce türetilmişse veri setine eklenmez.

Adım 6. İstenilen toplam veri sayısına ulaşana kadar 1. adımdan itibaren program bir döngü içinde tekrar çalıştırılır (bazı değerler için olası tüm verilerin sayısı, türetilmek istenen veri sayısından daha az olabilir. Örneğin 4 kategori için toplamı 10 olan ve tekrar etmeyen 1.000 tane veri bulmak imkânsızdır).

Adım 7. Son olarak türetilen veriler Python "Comma Separated Value (CSV)" kütüphanesi kullanılarak CSV formatında dosyalara yazdırılır.

$n = 10$ için 286, $n = 15$ için 815 ve $n \geq 20$ için biner farklı veri seti çalışmada kullanılmıştır. $a, b, c, ve\ d$'ye dayalı 23 farklı farklılık/uzaklık katsayısına ait formüller Tablo 2'de verilmiştir.

Ayrıca çalışmada dikkate alınan farklılık/uzaklık 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 IBM SPSS Statistics for Windows, Version 20.0. (Armonk, NY) paket programından yararlanılmıştır.

TABLO 2: Farklılık/uzaklık katsayıları.

No	Adı	Formül	Değer aralığı
1	Bray-Curtis/[5-8] Lance-Williams	$\frac{b+c}{2a+b+c}$	$< 0, +1 >$
2	Chord[5,6,9]	$\sqrt{2\left(1-\frac{a}{\sqrt{(a+b)(a+c)}}\right)}$	$< 0, +1{,}414 >$
3	Correlation[10]	$\frac{1}{2}-\frac{ad-bc}{2\sqrt{(a+b)(c+d)(a+c)(b+d)}}$	$< 0, +1 >$
4	Cosine[11]	$\frac{c}{\sqrt{ab}}$	$< 0, +\infty)$
5	Dice[11]	$\frac{2c}{a+b}$	$< 0, +\infty)$
6	Euclid I[5,6]	$\sqrt{b+c}$	$< 0, +\infty)$
7	Euclid II[11]	$\sqrt{a+b-2c}$	$< 0, +\infty)$
8	Hamming/[5,6,12] Canberra/Manhattan/ City-Block/Minkowski	$b+c$	$< 0, +\infty)$
9	Hellinger[5,6,13]	$2\sqrt{\left(1-\frac{a}{\sqrt{(a+b)(a+c)}}\right)}$	$< 0, +2 >$
10	Jaccard/[10] Needham	$\frac{b+c}{a+b+c}$	$< 0, +1 >$
11	Kulzinsky[10]	$\frac{b+c-a+n}{b+c+n}$	$< 0, +1 >$
12	Pattern difference[5,6] (yapı farkı)	$\frac{4bc}{n^2}$	$< 0, +1 >$
13	Rogers/[10] Tanimoto	$\frac{2b+2c}{a+d+2b+2c}$	$< 0, +1 >$
14	Russell/Rao[10]	$\frac{n-a}{n}$	$< 0, +1 >$
15	Shape difference[5,6] (şekil farkı)	$\frac{n(b+c)-(b-c)^2}{n^2}$	$< 0, +1 >$
16	Size difference[5,6] (boyut farkı)	$\frac{(b+c)^2}{n^2}$	$< 0, +1 >$
17	Soergel I[14]	$\frac{b+c}{b+c+d}$	$< 0, +1 >$
18	Soergel II[11]	$\frac{a+b-2c}{a+b-c}$	$(-\infty, +\infty)$
19	Sokal/[10] Michener	$\frac{n-a-d}{n}$	$< 0, +1 >$
20	Tanimoto[11]	$\frac{c}{a+b-c}$	$(-\infty, +\infty)$
21	Varyans[5,6]	$\frac{b+c}{4n}$	$< 0, +0{,}25 >$
23	Yule I[10]	$\frac{bc}{ad+bc}$	$< 0, +1 >$
22	Yule II[5,6]	$\frac{2bc}{ad+bc}$	$< 0, +2 >$

BULGULAR

Bu çalışmada, dikkate alınan farklılık/uzaklık katsayılarının çoğu iyi bilinmekte ve yaygın olarak kullanılmaktadır. Herhangi bir model kullanmadan kapsamlı ve varsayımsal olarak rastgele elde edilen uyum matrislerine dayalı 23 farklı farklılık/uzaklık katsayısına ait değerler elde edilmiş ve karşılaştırılmıştır. a, b, c ve d'nin özel durumlarda aldıkları değerler için elde edilen sonuçlar Tablo 3'te verilmiştir.

TABLO 3: a, b, c *ve* d'nin özel durumlarda aldıkları değerler için elde edilen sonuçlar.

	$a = n$	$b = n$	$c = n$	$d = n$	$a + d = n$	$b + c = n$	$a = b = c = d$
Bray-Curtis	0	1	1	NC	0	1	0,5
Chord	0	NC	NC	NC	0	1,414	1
Correlation	NC	NC	NC	NC	0	1	0,5
Cosine	NC	NC	NC	NC	NC	NC	1
Dice	0	0	NC	NC	0	C	1
Euclid I	0	C+	C+	0	0	C+	C+
Euclid II	C+	C+	NC	0	C	NC*	0
Hamming	0	n	n	0	0	n	$n/2$
Hellinger	0	NC	NC	NC	0	2	1,414
Jaccard-Needham	0	1	1	NC	0	1	0,667
Kulzinsky	0	1	1	1	C	1	0,833
Pattern difference (yapı farkı)	0	0	0	0	0	$0 \leq D \leq 1$	0,25
Rogers-Tanimoto	0	1	1	0	0	1	0,667
Russell-Rao	0	1	1	1	C	1	0,75
Shape difference (şekil farkı)	0	0	0	0	0	$0 \leq D \leq 1$	0,5
Size difference (boyut farkı)	0	1	1	0	0	1	0,25
Soergel I	NC	1	1	0	0	1	0,667
Soergel II	1	1	2	NC	1	C	0
Sokal-Michener	0	1	1	0	0	1	0,5
Tanimoto	0	0	-1	NC	0	C	1
Varyans	0	0,25	0,25	0	0	0,25	0,125
Yule I	NC	NC	NC	NC	0	1	0,5
Yule II	NC	NC	NC	NC	0	2	1

NC: Hesaplanamaz; D: Distance (uzaklık); C: n değerine bağlı olarak farklı değerler alır; C+: n değeri arttıkça değeri büyür; NC*: $b > c$ için hesaplanır ve n değerine bağlı olarak farklı değerler alır.

Tablo 3'te de görüldüğü üzere gözlem değerlerinin tamamının 2×2 tablodaki a gözesinde yer alması durumunda katsayıların önemli bir kısmı minimum değerini almakta bir kısmı ise hesaplanamamaktadır. Gözlem değerlerinin tamamının d gözesinde yer alması durumunda ise katsayıların önemli bir kısmı ya hesaplanamamakta ya da minimum değerini almaktadır. Katsayıların büyük bir kısmı $b + c = n$ olduğunda maksimum, $a + d = n$ olduğunda ise minimum değerlerine ulaşmaktadır. Gözelerde yer alan gözlem değerlerinin eşit olması durumunda ise katsayılar n sayısı ne olursa olsun sabit bir değer almaktadır. Hamming katsayısı dışındaki 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 4'te de görüldüğü üzere tüm n değerleri için Hamming katsayısı diğer katsayılardan ayrışmaktadır. $n \leq 30$ olduğu durumlarda Euclid I ve Euclid II katsayıları ayrı bir küme oluştursalar da özellikle $n \geq 35$ olması durumunda Hamming katsayısı dışındaki tüm katsayılar bir araya gelmektedirler.

TABLO 4: Hiyerarşik kümeleme analizi sonuçları.

n	Küme no	Yöntemler
10-30	1	Bray-Curtis, Chord, Correlation, Cosine, Dice, Hellinger, Jaccard-Needham, Kulzinsky, pattern difference (yapı farkı), Rogers-Tanimoto, Russell-Rao, shape difference (şekil farkı), size difference (boyut farkı), Soergel I, Soergel II, Sokal-Michener, Tanimoto, varyans, Yule I, Yule II
	2	Euclid I, Euclid II
	3	Hamming
≥ 35	1	Bray-Curtis, Chord, Correlation, Cosine, Dice, Euclid I, Euclid II Hellinger, Jaccard-Needham, Kulzinsky, pattern difference (yapı farkı), Rogers-Tanimoto, Russell-Rao, shape difference (şekil farkı), size difference (boyut farkı), Soergel I, Soergel II, Sokal-Michener, Tanimoto, varyans, Yule I, Yule II
	2	Hamming

TARTIŞMA

Farklılık, kesin bir matematiksel tanımı olmayan kavram olmasına rağmen farklılığın ölçülmesi belirli bir amaç için tasarlanabilecek bazı nicel ölçülere dayanmalıdır. Çok sayıda ölçü yayınlanmış olmasına rağmen farklılığı daha doğru, daha kapsamlı ve nesnel bir şekilde ölçmek için yeni ölçülerin geliştirilmesi bir öncelik olmaya devam etmektedir.[15,16] İki sonuçlu nitel veriler için çeşitli alanlarda çok sayıda farklılık/uzaklık ölçüleri önerilmiştir. İki sonuçlu veriler için olağan benzerlik ölçüleri yerine, olası asimetrileri hesaba katan ve dolayısıyla verileri değerlendirmek için farklı bir bakış açısı sağlayan farklılık/uzaklık ölçüleri kullanılabilir. Ancak verilerin kategorik olması farklılık/uzaklık ölçüleri ile ilgili ek problemler doğurmaktadır, çünkü genel olarak örtük bir farklılık/uzaklık fonksiyonu ile bir metrik uzayı tanımlamak mümkün değildir. Farklılık/uzaklık ölçüleri bunların yakınlıklarını belirtmek için kullanılabilir, ancak sonuçları şüphelidir. Farklılık/uzaklık ölçülerinin önemli bir kısmının 2 sonuçlu verilere uyarlanmadığına atıfta bulunularak genellikle uygun bir benzerlik ölçüsünün kullanılması daha çok önerilmektedir.

SONUÇ

Benzerlik kavramının tamamlayıcısı, farklılık kavramıdır. Ancak farklılık kavramı ile birlikte uzaklık terimi de sıklıkla kullanılmaktadır. 2×2 çapraz tablolar, benzerlik ölçüsüne alternatif olarak farklılık/uzaklık ölçüleri ile de değerlendirilebilir. Benzerlik katsayısının daha yüksek bir değeri, 2 farklı değişken arasında daha fazla ilişki olduğunu gösterirken, düşük bir değer 2 değişkenin farklı olduğunu gösterir. Farklılık/uzaklık katsayısı için ise yorum tam tersidir. İki sonuçlu veriler için uzaklık katsayılarının $0 \leq Uzaklık\ Katsayısı \leq 1$ aralığında değer alması beklenmesine rağmen tüm katsayılar için bu aralık geçerli değildir. Dikkate alınan 23 farklı katsayı içerisinden yalnızca 12 tanesi bu aralıkta değer almaktadır. Hiyerarşik kümeleme analizine göre uzaklık katsayılarının çoğu birbirine benzemektedir. Euclid I ve Euclid II katsayıları $n \leq 30$ değerleri için Hamming katsayısı ise $n \geq 30$ değerleri için diğer katsayılardan farklılık göstermektedir. Cosine, Euclid I, Euclid II ve Hamming katsayıları n sayısından etkilenmekte diğer katsayılar etkilenmemektedir. Dolayısıyla farklılık/uzaklık katsayılarının çoğunlukla örnek büyüklüğünden bağımsız oldukları belirlenmiştir. Genel olarak hemen hemen tüm katsayılara ait değerler, örnekler daha benzemez hâle geldikçe sabit bir minimumdan sabit bir maksimuma doğru artmaktadır. $b + c$ değeri arttıkça Cosine, Dice, Russell-Rao, Yule I ve Bray-Curtis katsayıları $n = 10\ ve\ 15$ için Hamming ve varyans katsayıları ise tüm n değerleri için farklılık/uzaklık ile doğrusal olarak artmaktadır. Rogers-Tanimoto katsayısı ise tüm n değerleri için doğrusal olmayan artış göstermektedir. Sokal-Michener katsayısı değer aralığının 0-1 olması ve farklılık/uzaklık artışı ile doğrusal olarak artmasından dolayı tüm katsayılar içerisinde öne çıkmaktadır. Literatürde farklılık/uzaklık ölçüleri ile ilgili teorik bilgi veren, ölçülerin karşılaştırıldığı yayın sayısı oldukça sınırlıdır. Çalışmalar daha çok uygulamaya yöneliktir. Dolayısıyla çalışmanın literatürdeki bu boşluğu dolduracağı düşünülmektedir.

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

Bu çalışma hazırlanırken tüm yazarlar eşit katkı sağlamıştır.

KAYNAKLAR

1. Cha SH. Comprehensive survey on distance/similarity measures between probability density functions. Int J Math Models Methods Appl Sci. 2007;4(1):300-7. [Link]
2. Le SQ, Ho TB. An association-based dissimilarity measure for categorical data. Pattern Recognit Lett. 2005;26(16):2549-57. [Crossref]
3. Willett P, Barnard JM, Downs GM. Chemical similarity searching. J Chem Inf Comput Sci. 1998;38(6):983-96. [Crossref]
4. Ballabio D, Todeschini R, Consonni V. Distances and other dissimilarity measures in chemometrics. In: Meyers RA, ed. Encyclopedia of Analytical Chemistry. 1st ed. New York: John Wiley & Sons; 2015. p.1-34. [Crossref]
5. Choi SS, Cha SH, Tappert CC. A survey of binary similarity and distance measures. J Syst Cybern Inf. 2010;8(1):43-8. [Link]
6. 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]
7. Bray JR, Curtis JT. An ordination of upland forest communities of Southern Wisconsin. Ecol Monogr. 1957;27(4):325-49. [Crossref]
8. Lance GN, Williams WT. A general theory of classificatory sorting strategies II. clustering systems. Comput J. 1967;10(3):271-7. [Crossref]
9. Orlóci L. An agglomerative method for classification of plant communities. J Ecol. 1967;55(1):193-206. [Crossref]
10. Zhang B, Srihari SN. Binary vector dissimilarities for handwriting identification. In Proceedings SPIE 5010, Document Recognition and Retrieval X, Jan 20-24, Santa Clara, California, USA: 2003. p.155-66. [Crossref]
11. Bajusz D, Rácz A, Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J Cheminform. 2015;7:20. [Crossref] [PubMed] [PMC]
12. Hamming RW. Error detecting and error correcting codes. The Bell System Technical Journal. 1950;29(2):147-60. [Crossref]
13. Hellinger E. Neue Begründung der Theorie quadratischer Formen von unendlichvielen veränderlichen. Journal für die reine und angewandte Mathematik (in German). 1909;136:210-71. [Crossref]
14. Monev V. Introduction to similarity searching in chemistry. MATCH-Commun Math Co. 2004;51(51):7-38. [Link]
15. Chao A, Chazdon RL, Colwell RK, Shen TJ. A new statistical approach for assessing similarity of species composition with incidence and abundance data. Ecol Lett. 2005;8(2):148-59. [Crossref]
16. Hao M, Corral-Rivas JJ, González-Elizondo MS, Ganeshaiah KN, Nava-Miranda MG, Zhang C, et al. Assessing biological dissimilarities between five forest communities. For Ecosyst. 2019;6(30). [Crossref]