ISSN 1870-4069

Generalidades para funciones de correlación entre distribuciones de probabilidad

Maria Elena Ensastegui-Ortega, Ildar Batyrshin, Alexander Gelbukh

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
México

{elena.ensastegui, batyr1}@gmail.com,
gelbukh@gelbukh.com

Resumen. El artículo presenta funciones de correlación basadas en similitudes y disimilitud entre distribuciones de probabilidad. La medidas de similitud y distancia que son usadas comparan dos distribuciones de probabilidad o tuplas de números reales. Se exploran funciones de distribución de probabilidad, medidas de distancias/similitud que cumplen con las propiedades de función de similitud o disimilitud. A partir de estas nuevas funciones de similitud o disimilitud se construyen nuevas funciones de correlación. También se da un bosquejo general de lo que son las funciones de similitud, disimilitud y correlación.

Palabras clave: Funciones de similitud o disimilitud, funciones complementarias, funciones de correlación.

Generalities for Correlation Functions between Probability Distributions

Abstract. The paper presents correlation functions based on similarity and dissimilarity between probability distributions. The similarity and distance measures that are used compare two probability distributions or tuples of real numbers. Probability distribution functions, measures of distance or similarity that meet the similarity or dissimilarity function properties are explored. From these new similarity or dissimilarity functions, new correlation functions are constructed. A general outline of what the similarity, dissimilarity, and correlation functions are is also given.

Keywords: Similarity or dissimilarity functions, complementary functions, correlation functions.

1. Introducción

Las aplicaciones a los algoritmos de Machine Learning y Minería de Datos han llevado al ser humano a un crecimiento acelerado en muchas áreas tanto del conocimiento como de la vida diaria, esto se debe a que los seres humanos cada día pueden usar herramientas que cumplan con tareas automatizadas dando a los seres humanos la opción de optimizar su tiempo.

Las medidas de similitud y correlación se utilizan en la recuperación de información, clasificación de datos, aprendizaje automático, análisis de relaciones y toma de decisiones en ecología, lingüística computacional, procesamiento de imágenes y señales, análisis de datos financieros, bioinformática y ciencias sociales.

Los algoritmos de aprendizaje no supervisado hacen uso de medidas de similitud para poder agrupar conjuntos de datos, estos algoritmos trabajan tanto con datos nominales como numéricos, estos datos se mide que tanto se relacionan entre sí, con medidas de similitud.

En este articulo, la segunda sección, se explican las propiedades de funciones de similitud y disimilitud, la tercera es sección una breve explicación sobre distribuciones de probabilidad, en la cuarta sección se exploran medidas de similitud como: Bhattacharyya, Coseno, Czekanowski, Ruzicka, Jaccard y Dice que también son funciones de distribuciones de probabilidad [2] se puede ver que cumplen que son funciones $S : \Omega \times \Omega \to [0; 1]$ s Ω si para todo x, y en Ω cumplen con las propiedades de simetría y reflexiva por tanto son funciones de similitud [1], a partir de estas se construyen nuevas funciones de correlación.

En la sección cinco se exploran medidas de distancias como: Sorensen, Wave Hedges, Soergel, Tanimoto, Jaccard y Dice, que también son funciones de distribuciones de probabilidad [2] se puede ver que cumplen que son funciones $S : \Omega \times \Omega \to [0; 1]$ s Ω si para todo x, y en Ω cumplen con las propiedades de simetría e irreflexiva por tanto son funciones de disimilitud [1] a partir de estas se construyen nuevas funciones de correlación.

En la sección seis, se muestra que algunas de estas funciones de similitud y disimilitud son complementarias. En la sección siete, se da un breve resumen de las funciones de correlación creadas. Por último en la sección siete se discuten las concluciones y trabajo a futuro.

2. Funciones de similitud, disimilitud y de correlacion

Los datos de los cuales podremos extraer conocimiento pueden venir de un conjunto diferente del vacío, sea Ω tal conjunto que se denominará un dominio universal o un conjunto subyacente en la definición de similitud. Ya que podemos considerar a Ω como un dominio específico para un tipo de datos considerado: el conjunto de todas las n-tuplas binarias, el conjunto de todos los vectores con valores en los reales de longitud n, el conjunto de imágenes u objetos considerados en algún problema, etc[1].

Una función $S : \Omega \times \Omega \to [0; 1]$ se llama una función de similitud en Ω si para todo x, y en Ω cumple las siguientes propiedades:

- Simetría: $S(x, y) = S(y, x)$ [5],
- Reflexiva: $S(x, x)$=1.

Una función D: $\Omega \times \Omega \to [0; 1]$ es una función de disimilitud en Ω si para todo x, y en Ω esta cumple las siguientes propiedades:

- Simetría: $D(x, y) = D(y, x)$,
- Irreflexiva: $D(x, x) = 0$.

Podemos decir que estas funciones son complementarias si para todo x, y en Ω se cumple que: $S(x, y) + D(x, y) = 1$.

Para funciones complementarias de similitud o disimilitud tenemos que:

$$S(x, y) = 1 - D(x, y), D(x, y) = 1 - S(x, y). \tag{1}$$

Una función A: $\Omega \times \Omega \to [0; 1]$ es una función de correlación en Ω si para todo x, y en Ω esta cumple las siguientes propiedades [5]:

- Simetría: $A(x, y) = A(y, x)$,
- Reflexiva: $A(x, x) = 1$,
- Negativa:$A(x, y) < 0$ para algún x, y en Ω.

Dichas funciones de correlación se denominarán funciones de correlación débiles si no satisfacen la propiedad de relación inversa la cual nos dice:

Proposición 1. Suponga que S y D son funciones de similitud y disimilitud en Ω de tal manera que para algunos x, y en Ω se cumple: $S(x, y) < D(x, y)$ y, entonces la función definida para todo (x, y) en Ω esta dada por:

$$A(x, y) = S(x, y) - D(x, y), \tag{2}$$

es una función de correlación. Si S y D son complementarios, entonces la función A será una función de correlación si para algún x, y en Ω se cumple: $S(x, y) <$,5.

La fórmula obtenida para A tiene una interpretación razonable: la correlación entre x e y es positiva si la similitud entre ellas es mayor que la disimilitud, y la correlación es negativa en caso contrario.

Si las funciones de similitud S y disimilitud D son complementarias, la función de correlación A definida por (1) se llama complementaria a S y D. Las funciones complementarias S, D y A se designarán como (S, D, A) y se denominarán tripleta de correlación . De la definición de las funciones complementarias disimilitud, similitud y de (2) se deduce que las funciones de similitud, disimilitud y correlación de la tripleta de correlación (S,D,A) pueden obtenerse una de otra para todo (x, y) en Ω como sigue [4]:

$$S(x, y) = 1 - D(x, y), D(x, y) = 1 - S(x, y), \tag{3}$$

$$A(x, y) = 2S(x, y) - 1, S(x, y) = \frac{1}{2}A(x, y) + 1. \tag{4}$$

$$A(x, y) = 1 - 2D(x, y), D(x, y) = \frac{1}{2}(1 - A(x, y)). \tag{5}$$

3. Distribuciones de probabilidad

Un espacio probabilístico formalmente es una tripleta (Ω, F, P) con (Ω, F) un espacio métrico y P una medida de probabilidad. Supongamos que existe $A \in F$ entonces podemos decir que $P(A)$ se le llama probabilidad de A [3].

Consideramos que P es una función que asigna valores en el intervalo (0,1) y cumple las siguientes propiedades: Sea w una variable aleatoria en el conjunto F [3]:

1. $P(w) \geq 0$,
2. $\sum_{w \in \Omega} p(w) = 1$.

Sea X un conjunto finito numerable y consideramos a la variable aleatoria w, decimos que $p_w(x) = P(C_x)$ con $\{C_x = y : w(y) = x\}$, es una distribución de probabilidad de la variable w.

En términos simples una función de distribución de probabilidad asigna evento que ocurre sobre la variable aleatoria la probabilidad de que dicho evento ocurra [6].

4. Construcción de funciones correlación con funciones de similitud

Sea $x = x_1, \ldots, x_n$ una distribución de probabilidad finita con $x_i \geq 0$ para todo $i = 1, \ldots, n$ y $\sum_{i=1}^{n} x_i = 1$. Sea $y = y_1, \ldots, y_n$ una distribución de probabilidad finita con $y_i \geq 0$ para todo $i = 1, \ldots, n$ y $\sum_{i=1}^{n} y_i = 1$. Definimos al coeficiente Bhattacharyya como sigue [5]:

$$S(x, y) = \sum_{i=1}^{n} \sqrt{x_i y_i}. \tag{6}$$

Podemos mostrar que la anterior similitud cumple las propiedades de simetría y reflexiva .

Ya que esta medida de similitud cumple con las propiedades para ser una funciń de similitud podemos concluir que es una función de similitud, por tanto podemos hacer uso de (4) para construir una función de correlación débil como sigue:

$$A(x, y) = 2 \sum_{i=1}^{n} \sqrt{x_i y_i} - 1. \tag{7}$$

La simititud coseno esta dada por la siguiente formula. Consideramos n-tuplas $x = (x_1, ..., x_n)$ y $y = (y_1, ..., y_n)$ con valores en los reales [4].

$$cos(x, y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}, \tag{8}$$

donde $x_i, y_i \geq 0$ $i = 1, ..., n$.

Es fácil notar que la similitud coseno cumple con las propiedades de simetría y reflexiva. Ya que es simétrica, reflexiva, es una función de similitud para construir una función de correlación débil consideramos la formula (4).

Entonces tenemos que la función de correlación débil de la función similitud coseno esta dada por:

$$A(x,y) = \frac{2\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} - 1. \tag{9}$$

La similitud Czekanowski compara dos distribuciones de probabilidad [2]. Sean P_i y Q_i dos distribuciones de probabilidad la similitud Czekanowski esta dada por:

$$S_{Cze} = \frac{2\sum_{i=1}^{d} min(P_i, Q_i)}{\sum_{i=1}^{d}(P_i + Q_i)}. \tag{10}$$

Se puede mostrar que la medida de similitud cumple con $S_{Cze} : \Omega \times \Omega \to [0;1]$, ademas de que es reflexiva y simétrica, entonces la similitud Czekanowski es una función de similitud, por las propiedades que cumplen las funciones de similitud por tanto podemos usar la ecuación (4) para construir una función de correlación débil para la función de similitud Czekanowski como sigue:

$$A_{Cze} = \frac{4\sum_{i=1}^{d} min(P_i, Q_i)}{\sum_{i=1}^{d}(P_i + Q_i)} - 1. \tag{11}$$

La similitud Ruzicka compara dos distribuciones de probabilidad [2]. Sean P_i y Q_i dos distribuciones de probabilidad la similitud Ruzicka esta dada por:

$$S_{Ruz} = \frac{\sum_{i=1}^{d} min(P_i, Q_i)}{\sum_{i=1}^{d} max(P_i, Q_i)}. \tag{12}$$

Se puede mostrar que la medida de similitud cumple con $S_{Cze} : \Omega \times \Omega \to [0;1]$, y cumple con las propiedades de simetría y reflexiva, por tanto podemos usar la ecuación (4) para construir una función de correlación débil para la similitud Ruzicka, como sigue:

$$A_{Ruz} = \frac{2\sum_{i=1}^{d} min(P_i, Q_i)}{\sum_{i=1}^{d} max(P_i, Q_i)} - 1. \tag{13}$$

La similitud Jaccard compara dos distribuciones de probabilidad [2]. Sean P_i y Q_i dos distribuciones de probabilidad la similitud Jaccard esta dada por:

$$S_{Jac} = \frac{\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i}. \tag{14}$$

Se puede mostrar que la medida de similitud cumple con $S_{Jac} : \Omega \times \Omega \to [0;1]$, y cumple con las propiedades de simetría y reflexiva, por tanto podemos usar

la ecuación (4) para construir una función de correlación débil para la similitud Jaccard, como sigue:

$$A_{Jac} = \frac{2\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i} - 1. \tag{15}$$

La similitud Dice compara dos distribuciones de probabilidad [2]. Sean P_i y Q_i dos distribuciones de probabilidad la similitud Dice esta dada por:

$$S_{Dice} = \frac{2\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2}. \tag{16}$$

Se puede mostrar que la similitud Dice tiene su dominio en el conjunto (0,1), y cumple con las propiedades de simetría y reflexiva, por tanto podemos usar la ecuación (3) para construir una función de correlación débil para la similitud Dice, como sigue:

$$A_{Dice} = \frac{4\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2} - 1. \tag{17}$$

5. Construcción de funciones correlación con funciones de disimilitud

La distancia Sorensen [2] puede ser usada para comparar dos distribuciones de probabilidad. Sean P_i y Q_i distribuciones de probabilidad, la distancia Sorencen esta dada por:

$$D_{sor} = \frac{\sum_{i=1}^{d} |P_i - Q_i|}{\sum_{i=1}^{d} P_i + Q_i}. \tag{18}$$

Se puede mostrar que esta distancia $D_{sor} : \Omega \times \Omega \rightarrow [0; 1]$, ademas de que es irreflexiva, y simétrica, entonces la distancia Sorencen es una función de disimilitud,por las propiedades que cumplen las funciones de disimilitud por tanto podemos usar la ecuación (5) para construir una función de correlación débil para la función de disimilitud Sorencen como sigue:

$$A_{sor} = 1 - \frac{4\sum_{i=1}^{d} |P_i - Q_i|}{\sum_{i=1}^{d} (P_i + Q_i)}. \tag{19}$$

La distancia Wave Hedges compara dos distribuciones de probabilidad [2]. Sean P_i y Q_i dos distribuciones de probabilidad, entonces la distancia Wave Hedges esta dada por:

$$d_{WH} = \sum_{i=1}^{d} \frac{1 - min(P_i, Q_i)}{max(P_i, Q_i)}. \tag{20}$$

Se puede mostrar que la distancia cumple con $d_{WH} : \Omega \times \Omega \rightarrow [0; 1]$) y cumple con las propiedades de simetría e irreflexiva, entonces podemos decir que es una

función de disimilitud, por tanto podemos usar la ecuación (5) para construir una función de correlación débil para la disimilitud Wave Hedges, como sigue:

$$A_{WH} = 1 - \sum_{i=1}^{d} \frac{2(1 - min(P_i, Q_i)}{max(P_i, Q_i))}. \tag{21}$$

La distancia Soergel compara dos distribuciones de probabilidad [2]. Sean P_i y Q_i dos distribuciones de probabilidad, la distancia Soergel esta dada por:

$$d_{Sg} = \frac{\sum_{i=1}^{d} |P_i - Q_i|}{\sum_{i=1}^{d} max(P_i, Q_i)}. \tag{22}$$

Se puede mostrar que esta distancia cumple con $D_{Sg} : \Omega \times \Omega \to [0; 1]$, ademas de que es irreflexiva, y simétrica, entonces la distancia Soergel es una función de disimilitud,por las propiedades que cumplen las funciones de disimilitud por tanto podemos usar la ecuación (5) para construir una función de correlación débil para la función de disimilitud Soergel como sigue:

$$A_{Sg} = 1 - \frac{2\sum_{i=1}^{d} |P_i - Q_i|}{\sum_{i=1}^{d} max(P_i, Q_i)}. \tag{23}$$

La distancia Tanimoto compara dos distribuciones de probabilidad [2]. Sean P_i y Q_i dos distribuciones de probabilidad, la distancia Tanimoto esta dada por:

$$D_{Tani} = \frac{\sum_{i=1}^{d} max(P_i, Q_i) - min(P_i, Q_i)}{\sum_{i=1}^{d} max(P_i, Q_i)}. \tag{24}$$

Se puede mostrar que esta distancia $D_{Tani} : \Omega \times \Omega \to [0; 1]$, ademas de que es irreflexiva, y simétrica, entonces la distancia Tanimoto es una función de disimilitud,por las propiedades que cumplen las funciones de disimilitud por tanto podemos usar la ecuación (5) para construir una función de correlación débil para la función de disimilitud Tanimoto como sigue:

$$A_{Tani} = 1 - \frac{2\sum_{i=1}^{d} max(P_i, Q_i) - min(P_i, Q_i)}{\sum_{i=1}^{d} max(P_i, Q_i)}. \tag{25}$$

La distancia Jaccard compara dos distribuciones de probabilidad [2]. Sean P_i y Q_i dos distribuciones de probabilidad la similitud Jaccard esta dada por:

$$D_{Jac} = \frac{\sum_{i=1}^{d} (P_i - Q_i)^2}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i}. \tag{26}$$

Se puede mostrar que la distancia cumple con $D_{Jac} : \Omega \times \Omega \to [0; 1]$, y cumple con las propiedades de simetría e irreflexiva, por tanto podemos usar la ecuación (5) para construir una función de correlación débil para la función de disimilitud Jaccard, como sigue:

$$A_{Jac} = 1 - \frac{2\sum_{i=1}^{d} (P_i - Q_i)^2}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i}. \tag{27}$$

La distancia Dice compara dos distribuciones de probabilidad [2]. Sean P_i y Q_i dos distribuciones de probabilidad la similitud Dice esta dada por:

$$D_{Dice} = \frac{\sum_{i=1}^{d}(P_i - Q_i)^2}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2}. \tag{28}$$

Se puede mostrar que la distancia cumple con $D_{Dice} : \Omega \times \Omega \to [0; 1]$, y cumple con las propiedades de simetría e irreflexiva, por tanto podemos usar la ecuación (5) para construir una función de correlación débil para la función de disimilitud Dice, como sigue:

$$A_{Dice} = 1 - \frac{\sum_{i=1}^{d}(P_i - Q_i)^2}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2}. \tag{29}$$

6. Funciones complementarias

Podemos notar que la función de disimilitud Sorensen puede expresarse de la siguiente manera [2]:

$$D_{sor} = 1 - S_{Cze} = \frac{\sum_{i=1}^{d}|P_i - Q_i|}{\sum_{i=1}^{d} P_i + Q_i}. \tag{30}$$

Lo que nos dice que por la ecuación (3) que la función de similitud Czekanowski y la función de disimilitud Sorensen son complementarias.

Se puede observar que, $S_{Jac} + D_{Jac} = 1$ [2] podemos concluir que son funciones complementarias por la ecuación (1). Sabemos que $S_{Dice} + D_{Dice} = 1$ [2] podemos concluir que son funciones complementarias por la ecuación (1).

Tabla 1. Funciones de similitud.

N	Funcion de Similitud	Funcion de Correlacion
(6)	$S(x,y) = \sum_{i=1}^{n}\sqrt{x_i y_i}$	$A(x,y) = 2\sum_{i=1}^{n}\sqrt{x_i y_i} - 1$
(8)	$cos(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}$	$A_{Cos}(x,y) = \frac{2\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} - 1$
(10)	$S_{Cze} = \frac{2\sum_{i=1}^{d} min(P_i,Q_i)}{\sum_{i=1}^{d}(P_i+Q_i)}$	$A_{Cze} = \frac{4\sum_{i=1}^{d} min(P_i,Q_i)}{\sum_{i=1}^{d}(P_i+Q_i)} - 1$
(12)	$S_{Ruz} = \frac{\sum_{i=1}^{d} min(P_i,Q_i)}{\sum_{i=1}^{d} max(P_i,Q_i)}$	$A_{Ruz} = \frac{2\sum_{i=1}^{d} min(P_i,Q_i)}{\sum_{i=1}^{d} max(P_i,Q_i)} - 1$
(14)	$S_J = \frac{\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i}$	$A_J = \frac{2\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i} - 1$
(16)	$S_{Dice} = \frac{2\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2}$	$A_{Dice} = \frac{4\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2} - 1$

Tabla 2. Funciones de disimilitud.

N	Funcion de Disimilitud	Funcion de Correlacion
(18)	$D_{sor} = \frac{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert}{\sum_{i=1}^{d} P_i + Q_i}$	$A_{sor} = 1 - \frac{4\sum_{i=1}^{d} \lvert P_i - Q_i \rvert}{\sum_{i=1}^{d} (P_i + Q_i)}$
(20)	$d_{WH} = \sum_{i=1}^{d} \frac{1 - min(P_i, Q_i)}{max(P_i, Q_i)}$	$A_{WH} = 1 - \sum_{i=1}^{d} \frac{2(1 - min(P_i, Q_i)}{max(P_i, Q_i))}$
(22)	$d_{Sg} = \frac{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert}{\sum_{i=1}^{d} max(P_i, Q_i)}$	$A_{Sg} = 1 - \frac{2\sum_{i=1}^{d} \lvert P_i - Q_i \rvert}{\sum_{i=1}^{d} max(P_i, Q_i)}$
(24)	$D_{Ta} = \frac{\sum_{i=1}^{d} max(P_i, Q_i) - min(P_i, Q_i)}{\sum_{i=1}^{d} max(P_i, Q_i)}$	$A_{Ta} = 1 - \frac{2\sum_{i=1}^{d} max(P_i, Q_i) - min(P_i, Q_i)}{\sum_{i=1}^{d} max(P_i, Q_i)}$
(26)	$D_J = \frac{\sum_{i=1}^{d} (P_i - Q_i)^2}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i}$	$A_J = 1 - \frac{2\sum_{i=1}^{d} (P_i - Q_i)^2}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i}$
(28)	$D_{Dice} = \frac{\sum_{i=1}^{d} (P_i - Q_i)^2}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2}$	$A_{Dice} = 1 - \frac{\sum_{i=1}^{d} (P_i - Q_i)^2}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2}$

7. Resultados

Por último en la Tabla 1 muestra las funciones de correlación creadas a partir de las funciones de similitud, mientras que la Tabla 2 muestra las funciones de correlación creadas a partir de las funciones de disimilitud.

8. Conclusiones y trabajo a futuro

En efecto se crearon nuevas funciones de correlación que cumplen con la propiedad de correlación, de medidas de similitud y distancias de distribuciones de probabilidad, que cumplían con las características para ser funciones de similitud o en su caso funciones de disimilitud por tanto era posible construir su función de correlación, también pudimos mostrar que algunas de ellas eran funciones de similitud complementarias.

Como trabajo a futuro se planea crear nuevas funciones de correlación que cumplan con la propiedad fuerte de correlación. Ademas de encontrar conjuntos de datos para los cuales se pueden aplicar y medir su precisión.

Referencias

1. Batyrshin, I.: Data Science: Similarity, Dissimilarity and Correlation Functions. (2019)
2. Sung-Hyuk Cha: Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. (2007)
3. Koralov, L., Sinai, Ya.: Theory of Probability and Random Processes. (2007)
4. Batyrshin, I.: Towards a general theory of similarity and association measures: similarity, dissimilarity and correlation functions. Journal of Intelligent and Fuzzy Systems, vol. 36, no. 4, pp. 2977–3004 (2019)
5. Batyrshin, I.: Constructing correlation coefficients from similarity and dissimilarity functions. Acta Polytechnica Hungarica, 16(10), pp. 191–204 (2019)

6. Díaz Mata, A.: Estadística aplicada a la administración y economía. México, McGraw Hill (2013)