Análisis de propiedades de medidas de similitud con atributos binarios

Iván Ramírez, Ildar Batyrshin

Instituto Politécnico Nacional, Centro de Investigación en Computación,
Ciudad de México, D.F.

{ramirez.alvarez.ipn, batyr1}@gmail.com

Resumen. El presente trabajo propone extender la visión basada en teoría de conjuntos sobre medidas de similitud en objetos con atributos binarios y que también pueden ser presentados en forma de tabla de contingencia. Se analiza las propiedades de diferentes medidas de similitud lo cual permite clasificarlas.

Palabras clave: Teoría de conjuntos, medidas de similitud, clasificación.

Analysis of Properties of Similarity Metrics with Binary Attributes

Abstract. We propose to extend an approach based on the set theory for similarity metrics with binary attributes that can be presented as a contingence table. We analyze different properties of similarity that allow us to classify them.

Keywords: Set theory, similarity metrics, classification.

1. Introducción

El análisis de medidas de similitud con atributos binarios y 2x2 tablas de contingencia ha recibido considerable atención durante muchos años en varios trabajos de minería de datos y reconocimiento de patrones [1,4-20]. Hasta este momento se han desarrollado decenas de medidas de similitud. Dependiendo del campo u objeto de estudio [6, 4], aplicación, estructura o la forma con que se trata a los atributos, se intenta seleccionar la medida más adecuada lo cual impactaría directamente en la precisión de los resultados obtenidos y además en la reducción de tiempo y proceso. Dicha tarea no es nada trivial puesto que en su estudio se consideran diferentes aspectos tales como cumplir con ciertas propiedades matemáticas [2] o que tan correlacionadas están dos medidas entre sí [5].

De manera informal una medida de similitud es una función cuyo valor real cuantifica la semejanza entre dos objetos. Esta es utilizada para medir hasta qué punto dos objetos, de acuerdo con los valores de sus atributos (características), son similares.

Es importante considerar los diferentes puntos de vista desde los cuales se analiza y mide la similitud entre instancias además de la forma utilizada por la medida en cuestión, por ejemplo medidas basadas en posibilidad y probabilidad [9, 21] que evalúan la generalidad y confiabilidad de ciertas reglas. Otra forma de medir similitud entre instancias es utilizando objetos cuyos atributos solo se encuentran en $\{0, 1\}$. Estos valores determinan la ausencia o presencia de una característica, pero más en general, una medida de similitud binaria estima las relaciones por las cuales dos objetos están siendo agrupados (relaciones taxonómicas).

Por otra parte en ocasiones también se consideran restricciones como "normalización" lo que obliga a que la medida se encuentre entre un intervalo dado, que por lo general se encuentra en [0, 1] o para ciertas medidas en [-1, 1] en donde inclusive en varios estudios no hacen explicita la diferencia de si una medida en cuestión se encuentra en los intervalos anteriores [7].

En este artículo son estudiadas diferentes propiedades que una medida de similitud deberá satisfacer. Se analizan aquellas que en la literatura se consideran fundamentales, tales como reflexividad y simetría; además de otras propuestas en [2, 3] como reflexividad estricta, débil similitud de reflexiones, cancelación de las reflexiones y no similitud de reflexiones.

Este trabajo propone una metodología para analizar y comparar diferentes medias de similitud basado en su cumplimiento de propiedades importantes para clasificar estas medidas y generar nuevas medias de similitud y de asociación con métodos considerados en [2,3]. Esta metodología está basada en la representación de medidas de similitud en términos de teoría de conjuntos que da un método sencillo de investigar propiedades de estas medidas. Se propone un agrupamiento de medidas basado en lo anterior. Las medidas utilizadas en este trabajo son más populares de área de reconocimiento de patrones pero son solo algunas de las mostradas en [7], sin embargo para trabajo futuro este estudio contempla ser extendido para todas estas medidas.

El resto de este trabajo está organizado de la siguiente manera: en la sección 2, se da la caracterización de las medidas de similitud utilizando conjuntos. En la sección 3 se realiza el estudio de las propiedades propuestas. En sección 4 se propone un agrupamiento basado en las propiedades estudiadas y por último en la sección 5 se exponen las conclusiones obtenidas.

2. Representación medidas de similitud en conjuntos

Sea $X = \{x_1, \dots, x_n\}$ el conjunto de atributos con valores binarios. Un objeto es un conjunto de atributos que este posee. Entonces sean $A \subseteq X$ y $B \subseteq X$ dos objetos que están definidos como conjuntos de atributos sobre X. Las medidas de similitud para datos binarios pueden ser expresadas como funciones de cuatro cantidades tales como lo presentado en Tabla 1:

a es número de atributos en común por los dos objetos A y B;

b es número de atributos en A pero no en B;

c es número de atributos en B pero no en A;

d es número de atributos que no se encuentran en ninguno de los dos A y B.

Tabla 1. Tabla de contingencia 2x2

	$\boldsymbol{B}$	$\bar{\boldsymbol{B}}$
A	a	b
$\bar{A}$	c	d

La misma tabla de contingencia aparece cuando A y B son variables binarias como $A = tiene\ gripa$ y B $= tiene\ temperatura\ alta$, en este caso el resultado de la muestra de $n = a + b + c + d$ personas está dado por:

a número de personas que poseen los atributos A y B;

b número de personas que poseen A pero no B;

c número de personas que poseen B pero no A;|

d número de personas que no poseen ninguno de los dos A y B.

En el siguiente texto para definir la caracterización de las medidas de similitud se utilizara la primera interpretación de Tabla 1. También esta tabla es posible escribirla en términos de conjuntos como en Tabla 2.

Tabla 2. Tabla de contingencia para los conjuntos A y B

	$\boldsymbol{B}$	$\bar{\boldsymbol{B}}$
A	$\lvert A \cap B \rvert$	$\lvert A \cap \bar{B} \rvert$
$\bar{A}$	$\lvert \bar{A} \cap B \rvert$	$\lvert \bar{A} \cap \bar{B} \rvert$

Lo anterior es importante debido a que existen diferentes caracterizaciones de esta tabla dependiendo del enfoque analizado [9, 15, 18]. Una vez establecido lo anterior se puede también reescribir aquellas medidas de similitud que trabajen con objetos binarizados. La Tabla 3 muestra las medidas más populares en las tareas de clasificación y reconocimiento de patrones para medir semejanza entre objetos cuyos valores de atributos son binarios con su correspondiente forma en conjuntos [7].

Como se sabe, un conjunto es una colección de objetos no ordenados, donde los elementos están separados por comas dentro de símbolos de llaves. No están ordenados debido a que $\{a, b\} = \{b, a\}$. Por lo tanto una medida de similitud $S(A, B)$ donde A y B son dos conjuntos también cumple con que: la similitud es grande cuando A y B están cerca, la similitud es pequeña cuando estos están lejos y normalmente es igual a 1 cuando estos son iguales (propiedad de reflexividad), y se encuentran entre el intervalo $[0, 1]$.

Lo anterior muestra que algunas medidas son idénticas entre sí a la hora de realizar la medición con sus valores, tal es el caso de las medidas de Dice y Czekanowski que tratan de la misma forma a sus variables. Además de Sokal y Michener se puede inferir que $|A \cap B| + |\bar{A} \cap \bar{B}| = \left|\overline{A \oplus B}\right|$, $|A \cap B| + |\bar{A} \cap B| + |A \cap \bar{B}| + |\bar{A} \cap \bar{B}| = |U|$ y $|A \cap B| + |\bar{A} \cap B| + |A \cap \bar{B}| = |A \cup B|$ para de esta manera aclarar su uso.

3. Propiedades estudiadas

Formalmente una medida de similitud está definida como una función $S: X \times X \rightarrow [0,1]$ que cumple con algunas propiedades consideradas posteriormente. En [2] y [3] se propone que estas deben cumplir también con las siguientes propiedades además de simetría y reflexividad:

P1. $S(A,B) = S(B,A)$	(simetría),
P2. $S(A,A) = 1$	(reflexividad),
P3. $S(A,B) < S(A,A)$ si $B \neq A$	(reflexividad estricta),
P4. $S(A,\bar{A}) < 1$	(débil similitud de reflexiones),
P5. $S(\bar{A},\bar{B}) = S(A,B)$	(cancelación de las reflexiones),
P6. $S(A,\bar{A}) = 0$	(no similitud de reflexiones).

Tabla 3. Definición de algunas medidas de similitud para datos binarios y su correspondiente representación en términos de conjuntos

	Representación Binaria	Representación en Conjuntos
1	$S_{\text{JACCARD}} = \frac{a}{a+b+c}$	$\frac{\lvert A \cap B \rvert}{\lvert A \cup B \rvert}$
2	$S_{\text{DICE,CZEKANOWSKI}} = S_{\text{NEI \& LI}} = \frac{2a}{2a+b+c}$	$\frac{2\lvert A \cap B \rvert}{2\lvert A \cap B \rvert + \lvert \bar{A} \cap B \rvert + \lvert A \cap \bar{B} \rvert}$
3	$S_{\text{3W-JACCARD}} = \frac{3a}{3a+b+c}$	$\frac{3\lvert A \cap B \rvert}{3\lvert A \cap B \rvert + \lvert \bar{A} \cap B \rvert + \lvert A \cap \bar{B} \rvert}$
4	$S_{\text{SOKAL \& SNEATH-I}} = \frac{a}{a+2b+2c}$	$\frac{\lvert A \cap B \rvert}{\lvert A \cap B \rvert + 2\lvert \bar{A} \cap B \rvert + 2\lvert A \cap \bar{B} \rvert}$
5	$S_{\text{SOKAL \& MICHENER}} = \frac{a+d}{a+b+c+d}$	$\frac{\lvert \overline{A \oplus B} \rvert}{\lvert U \rvert} = \frac{\lvert A \cap B \rvert + \lvert \bar{A} \cap \bar{B} \rvert}{\lvert U \rvert}$
6	$S_{\text{SOKAL \& SNEATH-II}} = S_{\text{GOWER \& LEGENDRE}} = \frac{2(a+d)}{2a+b+c+2d}$	$\frac{2(\lvert A \cap B \rvert + \lvert \bar{A} \cap \bar{B} \rvert)}{2\lvert A \cap B \rvert + \lvert \bar{A} \cap B \rvert + \lvert A \cap \bar{B} \rvert + 2\lvert \bar{A} \cap \bar{B} \rvert}$
7	$S_{\text{ROGER \& TANIMOTO}} = \frac{a+d}{a+2(b+c)+d}$	$\frac{\lvert A \cap B \rvert + \lvert \bar{A} \cap \bar{B} \rvert}{\lvert A \cap B \rvert + 2(\lvert \bar{A} \cap B \rvert + \lvert A \cap \bar{B} \rvert) + \lvert \bar{A} \cap \bar{B} \rvert}$

	Representación Binaria	Representación en Conjuntos
8	$S_{FAITH} = \frac{a + 0.5d}{a + b + c + d}$	$\frac{\lvert A \cap B \rvert + 0.5\lvert \bar{A} \cap \bar{B} \rvert}{\lvert U \rvert}$
9	$S_{RUSSEL\ \&\ RAO} = \frac{a}{a + b + c + d}$	$\frac{\lvert A \cap B \rvert}{\lvert U \rvert}$
10	$S_{OCHAI-I} = \frac{a}{\sqrt{(a + b)(a + c)}}$	$\frac{\lvert A \cap B \rvert}{\sqrt{(\lvert A \cap B \rvert + \lvert \bar{A} \cap B \rvert)(\lvert A \cap B \rvert + \lvert A \cap \bar{B} \rvert)}}$

En la tabla 4, se muestra la verificación de las propiedades P1 – P6 sobre las medidas propuestas dentro de la tabla 3, donde (●) implica que cumple con la propiedad y (○) que no la cumple.

Tabla 4. Tabla que muestra la verificación de las propiedades P1 – P6.

		P1	P2	P3	P4	P5	P6
1	$S_{JACCARD}$	●	●	●	●	○	●
2	S_{DICE}, $S_{CZEKANOWSKI}$, $S_{NEI\ \&\ LI}$	●	●	●	●	○	●
3	$S_{3W\text{-}JACCARD}$	●	●	●	●	○	●
4	$S_{SOKAL\ \&\ SNEATH\text{-}I}$	●	●	●	●	○	●
5	$S_{SOKAL\ \&\ MICHENER}$	●	●	●	●	●	●
6	$S_{SOKAL\ \&\ SNEATH\text{-}II}$, $S_{GOWER\ \&\ LEGENDRE}$	●	●	●	●	●	●
7	$S_{ROGER\ \&\ TANIMOTO}$	●	●	●	●	●	●
8	S_{FAITH}	●	○	●	●	○	●
9	$S_{RUSSEL\ \&\ RAO}$	●	○	●	●	○	●
10	$S_{OCHAI\text{-}I}$	●	●	●	●	○	●

Mostrar todas las comprobaciones de estas propiedades para cada una de las medidas de similitud tomaría mucho espacio, sin embargo se muestran algunas comprobaciones realizadas.

Para poder probar dichas propiedades, se retoman algunas identidades que resultan útiles, tales como:

$A - B = A \cap \bar{B}$ (diferencia entre conjuntos),
$A \cup A = A$, $A \cap A = A$ (leyes de idempotencia),
$\bar{\bar{A}} = A$ (ley de complementación),
$A \cup B = B \cup A$, $A \cap B = B \cap A$ (leyes de conmutatividad),

$\overline{A \cap B} = \bar{A} \cup \bar{B}$, $\overline{A \cup B} = \bar{A} \cap \bar{B}$ (leyes de De Morgan),
$A \cup \bar{A} = U$, $A \cap \bar{A} = \emptyset$ (leyes de complemento).

Ejemplo 1. La propiedad de simetría (P1) para Jaccard:

$$S(A,B) = \frac{|A \cap B|}{|A \cap B| + |\bar{A} \cap B| + |A \cap \bar{B}|}.$$

Se comprueba remplazando A por B respectivamente según lo establecido en la función S, dada por:

$$S(B,A) = \frac{|B \cap A|}{|B \cap A| + |\bar{B} \cap A| + |B \cap \bar{A}|}.$$

De lo anterior podemos observar que a través de las leyes de conmutatividad se obtiene:

$$S(A,B) = S(B,A)$$

es decir cumple con la propiedad.

Ejemplo 2. La propiedad de reflexividad (P2) para Faith:

$$S(A,B) = \frac{|A \cap B| + 0.5|\bar{A} \cap \bar{B}|}{|U|}.$$

Se comprueba de la siguiente forma:

$$S(A,A) = \frac{|A \cap A| + 0.5|\bar{A} \cap \bar{A}|}{|U|}.$$

Utilizando las identidades de leyes de complementos y de idempotencia es posible ver que $|A \cap A| = |A|$, $|\bar{A} \cap \bar{A}| = |\bar{A}|$ y $|\bar{A} \cap A| = 0$, $|A \cap \bar{A}| = 0$ respectivamente para los términos $|\bar{A} \cap B|$ y $|A \cap \bar{B}|$ en $|U|$. Por lo tanto se obtiene:

$$S(A,A) = \frac{|A| + 0.5|\bar{A}|}{|A| + |\bar{A}|}.$$

De ahí que para esta medida se tiene:

$$S(A,A) \neq 1,$$

es decir no cumple con la propiedad.

Ejemplo 3. La propiedad de no similitud de reflexiones (P6) para Russel & Rao:

$$S(A,\bar{A}) = \frac{|A \cap \bar{A}|}{|U|}.$$

Inmediatamente por ley de complemento se tiene que $|\mathrm{A} \cap \overline{\mathrm{A}}| = |\emptyset| = 0$ y por lo tanto:

$$S(A, \bar{A}) = 0,$$

es decir, cumple con la propiedad.

4. Agrupamiento por propiedades

Basándose en la tabla 4 entonces se puede agrupar en diferentes clases a estas medidas. Dichas clases pueden ser parcialmente ordenadas de tal manera que una clase contiene a otra. Las clases quedan definidas por las clases $C1, C2$ y $C3$ respectivamente y cada una de ellas está dada por:

- $C1 = \{P1 - P6\}$ y contiene a las medidas Sokal & Michener, Sokal & Sneath-II, Gower & Legendre y Roger & Tanimoto, con números 5 – 7.
- $C2 = \{P1 - P4, P6\}$ y contiene a las medidas Jaccard, Dice, Czekanowski, Nei & Li, 3W-Jaccard, Sokal & Sneath-I y Ochai-I o basándose en la numeración de la tabla 4, con números 1 – 4 y 10.
- $C3 = \{P1, P3, P4, P6\}$ y contiene a las medidas Faith y Russel & Rao o las con números 8 y 9.

Lo anterior también puede verse utilizando una notación conjuntista de inclusión donde:

$$C3 \subseteq C2 \subseteq C1.$$

5. Conclusiones

Este artículo presenta un estudio en relación a medidas de similitud populares en diferentes áreas como clasificación y reconocimiento de patrones, utilizando y extendiendo el punto de vista de conjuntos lo que permite analizar a estas medidas basándose en las propiedades propuestas P1 – P6. Este análisis permite ver que estas medidas pueden ser agrupadas en diferentes clases $C1, C2, C3$. Se puede mencionar que se encontró que una medida bastante popular como Russel & Rao que pertenece a clase $C3$ no cumple con propiedad de reflexividad P2 que normalmente está considerada como obligatoria para medidas de similitud; por esto el uso de esta medida en tareas de clasificación y reconocimiento de patrones es cuestionable. Se debe mencionar que las medidas de la clase $C1$ en comparación con las de la clase $C2$ cumplen una importante propiedad como lo es P5 y debido a esto resultarían más útiles en algunas tareas de análisis de datos.

Para trabajo futuro, a partir de las propiedades estudiadas puede extenderse continuando con el análisis sobre la mayoría de medidas de similitud, disimilitud y asociación conocidas en varios estudios [7, 9, 18] para establecer sus propiedades y clasificarlas dependiendo de lo anterior y generar nuevas medidas que tienen propiedades deseables.

Agradecimientos. El presente trabajo fue apoyado por el proyecto SIP 20162204 y BEIFI del IPN.

Referencias

1. Batagelj, V., Bren M.: Comparing resemblance measures. Journal of classification, Vol. 12, No. 1, pp. 73–90 (1995)
2. Batyrshin, I.Z.: Association measures on [0, 1]. Journal of Intelligent and Fuzzy Systems, Vol. 29, No. 3, pp. 1011–1020 (2015)
3. Batyrshin, I.Z.: On definition and construction of association measures. Journal of Intelligent & Fuzzy Systems, Vol. 29, No. 6, pp. 2319–2326 (2015)
4. Cha, S.H., Srihari, S.N.: On measuring the distance between histograms. Pattern Recognition, Vol. 35, No. 6, pp. 1355–1370 (2002)
5. Cha, S.H.: Comprehensive survey on distance/similarity measures between probability density functions. City, Vol. 1, No. 2, p. 1 (2007)
6. Cha, S.H.: Taxonomy of nominal type histogram distance measures. City, Vol. 1, No. 2, p. 1 (2008)
7. Choi, S.S., Cha, S.H., Tappert, C.C.: A survey of binary similarity and distance measures. Journal of Systemics. Cybernetics and Informatics, Vol. 8, No. 1, pp. 43–48 (2010)
8. Dunn, G., Everitt, B.S.: An introduction to mathematical taxonomy. Courier Corporation, (2004)
9. Geng, L., Hamilton, H.J.: Interestingness measures for data mining: A survey. ACM Computing Surveys (CSUR), Vol. 38, No. 3, p. 9 (2006)
10. Gower, J.C., Legendre, P.: Metric and Euclidean properties of dissimilarity coefficients. Journal of classification, Vol. 3, No. 1, pp. 5–48 (1986)
11. Guillet, F., Hamilton, H.J. (Eds.): Quality measures in data mining (43). Springer, (2007)
12. Hinkin, T.R.: A brief tutorial on the development of measures for use in survey questionnaires. Organizational research methods, Vol. 1 No. 1, pp. 104–121 (1998)
13. Jalali-Heravi, M., Zaïane, O.R.: A study on interestingness measures for associative classifiers. ACM Symposium on Applied Computing, pp. 1039–1046 (2010)
14. Lerman, I.C.: Les bases de la classification automatique. Gauthier-Villars, (1970).
15. Lesot, M.J., Rifqi, M., Benhadda, H.: Similarity measures for binary and numerical data: a survey. International Journal of Knowledge Engineering and Soft Data Paradigms, Vol. 1, No. 1, pp. 63–84 (2008)
16. Li, Y., Qin, K., He, X.: Some new approaches to constructing similarity measures. Fuzzy Sets and Systems, Vol. 234, pp. 46–60 (2014)
17. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to data mining (1). Boston: Pearson Addison Wesley (2006)
18. Tan, P., Kumar, V., Srivastava, J.: Selecting the Right Interestingness Measure for Association Patterns. ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (2002)
19. Veal, L.R.: Syntactic Measures and Rated Quality in the Writing of Young Children. Studies in Language Education, Vol. 8 (1974)
20. Vo, B., Le, B.: Interestingness measures for association rules: Combination between lattice and hash tables. Expert Systems with Applications, Vol. 38, No. 9, pp. 11630–11640 (2011)
21. Zadeh, L.A.: A note on similarity-based definitions of possibility and probability. Information Sciences, Vol. 267, pp. 334–336 (2014)