RNPS No. 2142
ISSN 2072-6287
Versión Digital

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

REPORTE TÉCNICO
Reconocimiento de Patrones

SERIE AZUL

Similaridad vs. distancia

José Ruiz-Shulcloper,
Jesús Ariel Carrasco-Ochoa y
José Francisco Martínez-Trinidad

RT_068 **enero 2015**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

RNPS No. 2142
ISSN 2072-6287
Versión Digital

SERIE AZUL

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

REPORTE TÉCNICO

Reconocimiento de Patrones

Similaridad vs. distancia

José Ruiz-Shulcloper,
Jesús Ariel Carrasco-Ochoa,
José Francisco Martínez-Trinidad

RT_068 **enero 2015**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

Similaridad vs. distancia

José Ruiz-Shulcloper[1], Jesús Ariel Carrasco-Ochoa[2] y José Francisco Martínez-Trinidad[2]

[1] Equipo de Reconocimiento de Patrones,
Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV),
La Habana, Cuba
jshulcloper@cenatav.co.cu
[2] Coordinación de Ciencias Computacionales,
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE),
Puebla, México
{ariel, fmartine}@ccc.inaoep.mx

Resumen. En el trabajo se exponen las diferencias fundamentales entre el concepto de similaridad y el de distancia y el papel que éstos desempeñan en la solución de problemas reales de Reconocimiento de Patrones. Se hace un análisis del empleo, en muchas ocasiones inadecuado, de herramientas basadas en la función de distancia o de la similaridad como el opuesto o el inverso de una función de distancia. Y se muestran extensiones de herramientas que fueron diseñadas para funciones de distancia al caso de funciones de similaridad, que no son necesariamente el opuesto o el inverso de una función de distancia, ni siquiera funciones simétricas,.

Palabras clave: similaridad, distancia, vecino más cercano, c-means, prototipos, reconocimiento lógico combinatorio de patrones.

Abstract. In this paper, the fundamental differences between distance function and similarity function are exposed. The role of these concepts in the solution of real world Pattern Recognition problems is emphasized. We analyze the use, many times inadequate, of tools based on distance function or similarity functions that are not necessarily the inverse or the opposite of a distance function. In addition, we show extensions of tools initially designed for distance functions to the case of similarity functions that are not the inverse or the opposite of a distance function, even non-symmetric functions.

Keywords: similarity, distance, nearest neighbor, c-means, prototypes, logical combinatorial pattern recognition.

1 Introducción

Muchos seres vivos somos capaces de reconocer objetos, situaciones, fenómenos y muchas otras cosas más, a partir de observaciones, datos, informaciones, conocimientos previamente adquiridos. ¿Cómo lo hacemos? Es algo que no ha sido establecido con toda precisión y fundamento aunque por supuesto existen muchas teorías al respecto. Una respuesta sólida a esta pregunta abriría la posibilidad, por ejemplo, de poder reconstruir este mismo proceso en dispositivos computacionales y brindaría mayores opciones para incrementar la potencialidad de los seres humanos. A pesar de que al respecto aún no hay esas respuestas contundentes, sí existen hipótesis muy plausibles sobre la base de las cuales se han

desarrollado modelos matemáticos y herramientas computacionales encaminadas a potenciar las capacidades de los seres humanos en muchas tareas de interés vital para el desarrollo de la humanidad. Una de estas ideas es asumir que la "*proximidad* entre objetos" es un punto de partida para el logro de esos propósitos [Pekalska, Duin, (2005)]. Sin embargo, este concepto, como el de "*cercanía*", también utilizado por varios autores, está íntimamente ligado al concepto de *distancia*; y esto es muy importante tenerlo en cuenta dada la prolífera cantidad de situaciones diversas en las que debemos realizar el proceso de reconocer objetos o fenómenos en los problemas del mundo real. Los "*parecidos*" o las "*diferencias*" entre objetos, fenómenos o eventos, no siempre los podemos ver en términos de una distancia, como veremos en algunos de los ejemplos que a continuación mencionaremos.

Por otro lado aparece un concepto, clave en todo el proceso de formalización de estas ideas: la *representación* de los objetos o fenómenos que queremos reconocer. Como ya sabemos, el estudio de los objetos para su reconocimiento mediante dispositivos computacionales se realiza no directamente sobre los mismos sino sobre sus representaciones en un cierto espacio. Espacio que es construido a partir de las variables o rasgos que describen a los objetos. Estas representaciones pueden referirse a la estructura propia de los objetos o a la descripción, digamos semántica, de los mismos. En este último tipo de representación es que concentraremos nuestro análisis, aunque mucho de lo que aquí analizaremos es extensible a los otros tipos de representación.

Para algunos autores, una de las cuestiones básicas en el reconocimiento de patrones es poder establecer las *diferencias* entre los objetos, fenómenos o eventos y por tanto para ellos la *disimilaridad* es el concepto sobre el que han enfocado sus desarrollos [Pekalska, Duin, (2005)]. Para estos autores sólo cuando las diferencias han sido observadas y caracterizadas, la similaridad empieza a desempeñar un rol. A partir de aquí llegan a la conclusión de que la disimilaridad es más fundamental que la similaridad, por lo que han enfocado el desarrollo de sus teorías sobre la base de dicho concepto. Sin embargo, estos autores no dan una definición formal de una función de disimilaridad. Lo que sí queda claro, como expresan los propios autores de estas ideas, es que la opción por las disimilaridades está avalada por el hecho que ellas pueden ser interpretadas como distancias en un espacio vectorial conveniente y en muchos casos ellas pueden ser intuitivamente más atractivas.

A diferencia de lo expuesto por los citados autores, existen muchos ejemplos prácticos en los que es la *analogía*, el *parecido*, la *similaridad*, entre los objetos, fenómenos o eventos, lo que proporciona la información básica para poder establecer el reconocimiento entre los mismos. Cuando nos enfrentamos a dos cuadros clínicos, expuestos sobre la base de la sintomatología y los signos observados en los pacientes, en muchos casos no son las diferencias, sino las semejanzas lo que nos permite establecer un diagnóstico. De hecho, las diferencias pueden ser muchísimas y no tienen valor clínico. Los llamados *factores de riesgo* en muchos padecimientos se establecen sobre la base de las semejanzas, no de las diferencias. Algo análogo ocurre con el establecimiento de zonas perspectivas para la existencia de yacimientos minerales, o para el cultivo en la agricultura, o para el establecimiento de un diagnóstico técnico del funcionamiento de un equipo, o el pronóstico meteorológico, o el modus operandi de acciones delictivas, o la conducta de grupos de clientes en mercados, o de turistas en ciertas ciudades y muchos ejemplos más donde es la similaridad la que resulta ser el factor fundamental. A este hecho se une otro de carácter básico y que es característico de las ciencias poco formalizadas (Medicina, Geociencias, Sociología, Psicología, Ecología, la Criminalística y otras): la *frecuencia* de aparición de las combinaciones de valores de esos rasgos.

Es por ello que varios autores coincidimos en la idea de que uno de los conceptos fundamentales del Reconocimiento de Patrones es la *analogía* entre objetos, el *parecido* entre las descripciones de los objetos [Hubert, L.J. (1973), Tversky, A. (1977), Moskalienskii, E.D. (1984), Moskalienskii, E.D., Chinaiev, Y.B. (1984), Goldfarb, L. (1985), Sato, M., Sato, Y. (1995), Rodríguez, A., Egenhofer, M. (2003, 2004), Schwering, A., Raubal, M. (2005)]. Muy en particular, este concepto, como mencionamos anteriormente, es clave en las llamadas ciencias poco formalizadas, que por demás tienen un gran impacto en la sociedad desde muchos puntos de vista: económico, político, social, etc.

La formalización del concepto de similaridad (o del de disimilaridad) es una tarea importante que ha sido comúnmente reducida al uso de las distancias.

Cuando el estudio se realiza en un espacio de representación métrico, lo usual es considerar la *proximidad* como sinónimo conceptual de la *analogía*. Mientras más parecidos (semejantes, análogos) son los objetos que se comparan, más cercanas deberán estar sus representaciones en el espacio que se considere. Y por supuesto que existen problemas del mundo real que satisfacen estas condiciones. Sin embargo, cuando no hablamos de espacios métricos, cuando no es una distancia lo que se debe emplear para comparar las descripciones de los objetos, es natural que esta idea deba ser ajustada a las nuevas circunstancias.

El procedimiento empleado durante mucho tiempo por muchos investigadores ha sido considerar esa analogía, parecido, similaridad, como el opuesto o el inverso de una función de distancia. Y esto es, en ocasiones, correcto, pero no siempre los problemas del mundo real responden a estas exigencias.

Para entender mejor estas ideas es importante recordar las principales diferencias y conexiones entre estos dos conceptos: *distancia* y *similaridad* (*semejanza*). Aunque a diferencia del concepto de distancia, en el caso de la similaridad (o disimilaridad) no existe una definición universalmente aceptada, lo que dificulta el análisis de estas diferencias.

2 Algunos conceptos básicos

¿Qué es una distancia? Existe una variedad de funciones de distancia (pseudo-distancia, semi-distancia, distancia, pre-distancia, ultra-distancia, etc.), o más bien de medidas de disimilaridad, en dependencia del cumplimiento de cierto conjunto de propiedades. Aunque a muchas de estas "variaciones" de distancia se les denomina usando el término disimilaridad, no se da una definición formal de esta última.

Siguiendo a [Goldfarb (1985)], se verán algunos de los conceptos mencionados:

Definición 1. Por una *pseudo-distancia* (*coeficiente de disimilaridad*) se entiende una función real no negativa π, definida sobre el producto cartesiano de un conjunto P, que satisface las siguientes dos propiedades:

a) $\forall p_1,p_2 \in P\ \pi(p_1,p_2)=\pi(p_2,p_1)$ (simetría)
b) $\forall p \in P\ \pi(p,p)=0$ (reflexividad)

Definición 2. Una pseudo-distancia se denomina *semi-distancia* si satisface además:

c) $\forall p_1,p_2 \in P\ \pi(p_1,p_2)=0 \Rightarrow p_1=p_2$ (reflexividad fuerte).

Observemos que esta propiedad implica que el único caso en que la semi-distancia se anula, es decir, que alcanza su mínimo valor, es cuando los objetos que se comparan son los mismos. Esta propiedad también está presente en el concepto de distancia.

Finalmente,

Definición 3. Una semi-distancia que satisface la siguiente condición:

d) $\forall p_1,p_2,p_3 \in P\ \pi(p_1,p_3) \leq \pi(p_1,p_2)+\pi(p_2,p_3)$ (desigualdad triangular)
se le denomina *función de distancia* o simplemente *distancia*.

A pesar de la claridad de estas definiciones no es extraño encontrarse en la literatura de Reconocimiento de Patrones y Minería de Datos, autores que hablan de *distancias no simétricas*, o de *distancias que no cumplen la desigualdad triangular* lo que son expresiones incorrectas. De hecho la literatura recoge, entre otras, una función muy conocida y empleada en esta área del conocimiento que

responde al nombre de *distancia de Mahalanobis*, que no es una distancia, pues no cumple la condición c) de la definición 2 [Xing et al. (2002), Scheirer et al. (2014)].

Incluso el propio Goldfarb ha considerado que la condición c) no es una condición necesaria en toda función distancia, ya que como él mismo dice, uno puede requerir de la posibilidad de que dos objetos sean completamente similares con respecto a la función de distancia seleccionada, sin ser idénticos. Esto es una clara contradicción con la definición de distancia y de lo que se trata en todo caso es de la necesidad de usar coeficientes de disimilaridad como el propio Goldfarb los denomina, es decir, emplear la función que realmente responda a los requisitos específicos del problema que se está modelando y no adaptar la realidad a los prerrequisitos del modelo que se desea emplear, por más famoso y popular que este sea, pues el precio que pudiera pagarse por ese hecho puede ser mayor que el de buscar la función adecuada. En otras palabras, realizar la modelación matemática correcta del problema [Ruiz-Shulcloper, et al. (2013)] que se requiere resolver es lo que se impone en estos casos.

Pero el problema no es de un mero formalismo matemático ni de una exquisitez de notación o denotación. Es que muchas de las herramientas matemáticas útiles para la solución de problemas, en los que esas funciones están involucradas, asumen la existencia de una distancia no de una pseudo-distancia u otra función relativa a la distancia, aunque ésta además satisfaga la desigualdad triangular. Esto es, en este caso, que se cumpla la condición c) de la definición 2. Es importante enfatizar que las violaciones no son en cuanto a las denominaciones, sino en cuanto a la suposición de que ciertas funciones, necesarias para resolver un problema concreto de Reconocimiento de Patrones, satisfagan o no las hipótesis sobre las que descansa la herramienta matemática que se aplique.

Lo frecuente en la literatura es asumir la definición de distancia en términos de las cuatro propiedades antes mencionadas, o aún más directo, asumir distancias ya muy conocidas como son la Euclidiana, Minkowski, y otras. Incluso la denominada distancia de Mahalanobis, que no es una distancia, como ya hemos mencionado.

En ese mismo trabajo de Goldfarb se puede apreciar que la función de distancia, en general, ha sido considerada en la práctica como un equivalente de la función de disimilaridad. A partir de esta consideración, la mayoría de los análisis en Reconocimiento de Patrones se han realizado en términos de la distancia y después de arribar a las conclusiones, relacionadas directamente con los problemas del mundo real, las funciones de similaridad han sido interpretadas en forma del *opuesto* o del *inverso* de la distancia. Y este hecho se puede explicar, por el poder y desarrollo de las herramientas matemáticas elaboradas sobre la base del concepto de distancia. Sólo que una distancia, como ya hemos mencionado, es un caso particular de una disimilaridad por lo que no se justifica siempre esta consideración.

Esto explica por qué a muchos autores se les hace muy natural, más bien conveniente, pensar que si se asume una función de distancia, se puede obtener una función de similaridad mediante el opuesto o el inverso de dicha función. Y es cierto, el inverso o el opuesto de una distancia es una similaridad pero no son las únicas funciones de similaridad que son necesarias en la solución de los problemas reales de Reconocimiento de Patrones y Minería de Datos. Y con esa afirmación sí se podría estar totalmente de acuerdo si se hablara en general de que una similaridad es el inverso o el opuesto de una disimilaridad. Pero qué es una similaridad y qué es una disimilaridad.

Es importante subrayar que existe una gran cantidad de problemas reales, en particular relacionados con imágenes y señales, entre otros, en los que trabajar con una función de distancia es adecuado. En otros problemas es posible incluso que se pueda trabajar con una función de distancia, aprovechando el arsenal matemático desarrollado para estas funciones y luego interpretar los resultados en términos del opuesto o el inverso, según sea el caso, sin que con esto se deforme la realidad ni se cometa ninguna violación metodológica. Sin embargo, también es necesario destacar que esto no siempre es correcto y que *existen problemas reales en los que no es adecuado tal proceder*.

Finalmente, se puede concluir que no es totalmente cierto que una función de similaridad sea el opuesto o el inverso de una distancia. Por ejemplo, baste considerar un problema en el que dos objetos son totalmente similares en términos de sus descripciones aunque no sean idénticos. Situación que es muy frecuente en la práctica del mundo real: muchos pacientes descritos en términos de sus signos y síntomas tienen exactamente el mismo cuadro clínico no siendo idénticos sus signos y/o síntomas. En

este ejemplo, dado que los pacientes no son idénticos, el valor de la distancia sería diferente de cero por lo que el valor que se obtendría con el opuesto de la distancia no reflejaría que son totalmente similares.

Asumiendo que se pudiera obviar la condición c) en la definición de distancia, eludiendo las dificultades antes mencionadas, es posible encontrar problemas reales en los que la similaridad entre las descripciones de los objetos se realiza en términos de *funciones que no son simétricas*. Ejemplos de esta situación se pueden encontrar en el área de las Ciencias de la Información, en la Psicología, en las Geociencias, etc.

En [Chen, et al. (1992); (1997)] los autores abordaron la solución de un problema de agrupamiento de documentos basados en conjuntos de palabras claves o descriptores, para la realización de un sistema de recuperación de información. En estas investigaciones, Chen y sus colaboradores consideraban de interés los vínculos entre dos vocabularios especializados con un 30% de solapamiento. Los términos fueron pesados usando el producto de la frecuencia de los términos y el inverso del producto de la frecuencia de los documentos y agruparon los documentos usando una función que no es simétrica en la que se penalizan a los términos que aparecen con mucha frecuencia en los documentos y fijaron el número máximo de documentos en 100.

La función de similaridad empleada fue denominada por los autores como *función cluster*, y viene dada por las siguientes expresiones

$$F(T_j,T_k)=\frac{\sum_{i=1}^{n} d_{ij}d_{ik}}{\sum_{i=1}^{n} d_{ij}}, \qquad F(T_k,T_j)=\frac{\sum_{i=1}^{n} d_{ij}d_{ik}}{\sum_{i=1}^{n} d_{ik}},$$

en la que n es el número total de documentos en la base de datos, T_k denota el k-ésimo descriptor, d_{ij} es un rasgo booleano que denota la presencia o ausencia del término T_j en el i-ésimo documento.

Ambas expresiones pueden ser pesadas por un factor que dependa de T_k en el primer caso y de T_j en el segundo.

Como se puede observar, esta función de similaridad no es simétrica y obviamente no puede ser el inverso o el opuesto de función distancia alguna.

En [Sato (1992); Sato, M., Sato, Y. (1995)] se muestran los trabajos realizados por Y. Sato y M. Sato, en el marco de investigaciones para agrupar conjuntos de personas con respecto a relaciones de afinidad entre los mismos. Aquí, ellos introducen tres modelos difusos de agrupamiento como una generalización de los modelos aditivos de [Shepard, Arabie (1979)], y posteriormente extienden esos modelos para poder procesar datos usando relaciones que no son simétricas. Por ejemplo, la similaridad entre dos objetos en uno de esos modelos viene dada por:

$$\tilde{s}_{ij}=\sum_{k=1}^{K}\sum_{l=1}^{K} w_{kl}\mu_{ik}\mu_{jl}\ ;$$

donde K es el número de agrupamientos, μ_{ik} denota el grado en el que el objeto O_i pertenece al agrupamiento k, y w_{kl} es el peso en términos de la similaridad, que no es simétrica, entre un par de agrupamientos. Puede ocurrir que para algunos objetos $\mu_{ik}\mu_{jl}=\mu_{jk}\mu_{il}$, pero esto no necesariamente ocurre para todos los objetos y para algunos objetos puede ocurrir que $\mu_{ik}\mu_{jl}\neq\mu_{jk}\mu_{il}$, pero de nuevo, no necesariamente para todos. Es decir, la función no es simétrica.

Y. Sato y M. Sato, en el citado trabajo de 1995, muestran un ejemplo de aplicación de estos modelos usando las relaciones humanas entre un conjunto de jóvenes de 16 años. Esos datos representaban el grado de preferencias o no entre los jóvenes.

A los trabajos mencionados anteriormente se unen otros en diversas áreas del conocimiento, en las cuales también aparecen situaciones en las que las funciones de similaridad no son simétricas (aunque erróneamente en muchos de estos trabajos se denominan asimétricas, lo cual implicaría que la simetría no se cumple para par de objetos alguno). Los trabajos del grupo de M. Egenhofer sobre comparación entre entidades geoespaciales y el empleo de medidas de similaridad semántica entre estas entidades [Rodríguez, Egenhofer (2003); (2004); Schwering, Raubal (2005)] es uno de los ejemplos a considerar.

En la literatura, las funciones de similaridad no han recibido la misma atención que las funciones de distancia. Se puede encontrar una definición sencilla de función de similaridad en el libro de Jain y Dubbes [Jain, Dubes (1998)] o una más formal en [Alba (1998); Martínez-Trinidad, et al. (2000)]. En la literatura en ruso existen varios trabajos que intentan introducir una formulación axiomática de este concepto [Kochetkov (1978); Voronin (1985)]. Tampoco es inusual encontrar definiciones que contradicen otras ya establecidas, por ejemplo en el libro Clustering [Xu, Wunsch (2009)].

Independientemente de la ausencia de un acuerdo acerca del concepto de similaridad, la práctica ha impuesto la necesidad de utilizar, para la comparación entre descripciones de objetos, funciones más flexibles que la función distancia. Además de los ejemplos anteriormente mencionados acerca de la no simetría de la similaridad, se pueden encontrar otros ejemplos prácticos en los que también aparecen funciones menos restrictivas que la distancia [Mahalanobis (1936); Sebestyen (1962); Bongard (1963); Gower (1966); (1967); (1971); (1977); Michalski (1969); Baskakova, Zhuravlev (1981); Ruiz-Shulcloper, Fuentes (1981); Douglas-De la Peña, Ruiz-Shulcloper (1983); Goldfarb (1985); López-Reyes, et al. (1988); Álvarez-Gómez, et al. (1992); Ruiz-Shulcloper, et al. (1992); Gómez-Herrera, et al. (1994); Ralambondrainy (1995); Ortíz-Posadas, et al. (1996); (1997a); (1997b); (1998a); (1998b); (1998c), (1999), (2001)].

3 Similaridades que no son el inverso o el opuesto de una distancia

En esta sección se introduce una definición general de similaridad (disimilaridad) que permite construir funciones de similaridad (disimilaridad) que no son el inverso o el opuesto de una distancia y que pueden ser definidas para comparar datos mezclados e incompletos.

Por una <u>*descripción mezclada e incompleta*</u> de un objeto $O \in U$ (universo) entenderemos un n-uplo, $I(O)=(x_1(O),...,x_n(O))$, de valores numéricos y no numéricos (nominales, ordinales, booleanos, k-valuados) con la posible inclusión de un símbolo especial ($*$ ó ?) para denotar la ausencia de valores, donde $x_i(O) \in M_i$; siendo M_i el conjunto de valores admisibles del rasgo x_i, $i=1,...,n$.

Se considerará $M_1 \text{x} \ldots \text{x} M_n$, el producto cartesiano de los conjuntos de valores admisibles de los rasgos de $R=\{x_1,\ldots,x_n\}$ en términos de los cuales se describirán los objetos del universo U, como el espacio donde se representan (*espacio de representación*) los objetos bajo estudio, de esta forma $I(O) \in M_1 \text{x} \ldots \text{x} M_n$. Sobre este producto cartesiano no se asume estructura algebraica o lógica alguna. Por comodidad en las notaciones se identifica a los objetos con sus descripciones y se denotan de igual manera $I(O)=O$. Sea $M=\{O_1,...,O_m\} \subseteq U$, un conjunto de objetos.

Para cada rasgo x_i ($i=1,...,n$), se asocia un criterio de comparación de sus valores $C_i: M_i \times M_i \longrightarrow L_i$ donde:

$C_i(x_i(O),x_i(O)) = \min_{y \in L_i} \{y\}$, si C_i es un <u>*criterio de comparación de disimilaridad*</u> entre los valores del rasgo x_i ó

$C_i(x_i(O),x_i(O)) = \max_{y \in L_i} \{y\}$, si C_i es un <u>*criterio de comparación de similaridad*</u>.

C_i es una evaluación del grado de similaridad (o disimilaridad) entre dos valores de un mismo rasgo x_i siendo L_i un conjunto totalmente ordenado, $i=1,\ldots,n$, no necesariamente un conjunto numérico.

Para cada par de objetos en U, se puede calcular una magnitud que evalúe la similaridad entre los mismos (entre sus descripciones).

Denominaremos *función de similaridad* (*semejanza*) y la denotaremos por Γ a una función

$$\Gamma: \bigcup_{\Omega \subseteq R} \left(M_{i_1} x ... x M_{i_s}\right)^2 \longrightarrow L,$$

donde L es un conjunto totalmente ordenado; $\Omega=\{ x_{i_1},..., x_{i_s} \} \subseteq R$; $M_{i_1} \times ... \times M_{i_s}$ el producto cartesiano de sus respectivos conjuntos de valores admisibles; $s \geq 1$ que satisface las siguientes dos condiciones.

A) Condición de concordancia con las evaluaciones parciales:

Sean $T_1,\ldots,T_s$ subconjuntos no vacíos disjuntos de R, $\lhd$ el orden total sobre L y $T=\bigcup_{i=1}^{s} T_i$, entonces tenemos que:

Si para todo $h=1,\ldots,s$ $\Gamma(I|_{T_h}(O_i), I|_{T_h}(O_j)) \lhd \Gamma(I|_{T_h}(O_f), I|_{T_h}(O_g))$, entonces

$$\Gamma(I|_{T}(O_i), I|_{T}(O_j)) \lhd \Gamma(I|_{T}(O_f), I|_{T}(O_g)),$$

donde $I|_{T_h}(O_i)$ denota la subdescripción de un objeto O_i en términos de los rasgos de T_h;

B) Condición de máxima similaridad:

Para toda subdescripción en $\bigcup_{T \subseteq R} \left(M_{i_1} x ... x M_{i_s}\right)$, para todo $i,j=1,\ldots,s$ se cumple:

a) $\forall O_i \in U \; \forall T \subseteq R \; \max_{O_j \in M} \{ \Gamma\left(I|_T(O_i), I|_T(O_j)\right)\} = \Gamma\left(I|_T(O_i), I|_T(O_i)\right)$.

b) $\forall T_i, T_j \subseteq R \; \forall O \in U \; \Gamma\left(I|_{T_i}(O), I|_{T_i}(O)\right) = \Gamma\left(I|_{T_j}(O), I|_{T_j}(O)\right)$.

c) $\forall T \subseteq R \; \forall O_j, O_i \in U \; \Gamma\left(I|_T(O_i), I|_T(O_i)\right) = \Gamma\left(I|_T(O_j), I|_T(O_j)\right)$.

La condición de máxima similaridad es la formalización de un aspecto intuitivo: un objeto no se debe parecer más a otro objeto que a sí mismo en cualquier subconjunto de rasgos en que se tomen las descripciones.

La condición de concordancia con las evaluaciones parciales garantiza de que haya coherencia entre evaluaciones parciales y la evaluación total, es decir, si dada una colección de subconjuntos disjuntos no vacíos de R cuya unión es un conjunto T y además en cada elemento de dicha colección la similaridad entre un determinado par de objetos es siempre menor que la similaridad entre otro par, entonces la similaridad con respecto a T debe conservar la misma relación en los pares de objetos analizados.

Toda restricción de Γ a cualquier subconjunto T de R, la denominaremos *función de similaridad parcial.*

Análogamente se puede definir una función de disimilaridad.

Como se puede observar en esta definición no se exige la condición de simetría, que respondería a la intuición de que el grado de analogía (similaridad) de un objeto respecto a un segundo debería ser la misma que del segundo al primero. Sin embargo, en la práctica se pueden encontrar criterios de analogía, en diferentes ramas del conocimiento, en especial en las ciencias poco formalizadas, donde esta propiedad no se cumple. Ejemplos de este tipo de funciones pueden encontrarse, ya lo mencionamos anteriormente, en los trabajos de [Sato (1992); Sato, M., Sato, Y. (1995)], [Rodríguez, Egenhofer (2003); (2004)], [Schwering, Raubal (2005)] entre otros.

Podemos decir en general que el conjunto sobre el que se define la función de similaridad ($M_1 x \ldots x M_n$) es simplemente un producto cartesiano de los conjuntos de valores admisibles de los rasgos en términos de los cuales se describen los objetos de U (el universo de objetos) al que no se le presupone alguna propiedad algebraica, topológica o lógica.

Esta definición nos permitirá la comparación entre descripciones de objetos; cuando $s<n$, éstas serán comparaciones entre subdescripciones de los objetos (descripciones parciales de los objetos). En ocasiones la función de similaridad total depende de funciones de similaridad parciales, digamos por caso: es una combinación lineal de ellas [Martínez-Trinidad, et al. (2000)].

En la literatura se han utilizado diversas funciones de similaridad para comparar subdescripciones mezcladas e incompletas de objetos. Por ejemplo, en [Ayaquica-Martínez, et al. (2006), Hernández-Rodríguez, et al. (2008)] se usa una función de disimilaridad basada en criterios de comparación de similaridad entre valores de rasgos, definida como:

$$D(O,O') = 1 - \frac{\sum_{i=1}^{n} C_i(x_i(O), x_i(O'))}{n},$$

donde $C_i(x_i(O),x_i(O'))$ es un criterio de comparación de similaridad entre valores del rasgo x_i, $i=1,\ldots,n$, en los objetos O y O', definido como:

Cuando el rasgo x_i es no numérico:

$$C_i(x_i(O), x_i(O')) = \begin{cases} 1 & si\ x_i(O) = x_i(O') \\ 0 & en\ otro\ caso \end{cases}.$$

Cuando el rasgo x_i es numérico:

$$C_i(x_i(O), x_i(O')) = \begin{cases} 1 & si\ |x_i(O) - x_i(O')| < \sigma_i \\ 0 & en\ otro\ caso \end{cases},$$

donde σ_i es la desviación estándar del rasgo x_i en M_i.

Para el manejo de la ausencia de información, se considera que:

$$C_i(x_i(O), x_i(O')) = 0, \qquad si\ x_i(O)\ o\ x_i(O')\ son\ valores\ faltantes.$$

4 Tipos de funciones de similaridad

Una clasificación de las funciones de similaridad introducida en [Alba Cabrera (1998)] permite disponer de manera sencilla y práctica de estas funciones con fines de su utilización en la solución de problemas concretos. Muy en particular por la variedad de situaciones diversas que nos encontramos en la práctica real de los problemas de reconocimiento de patrones, no es factible tener un listado de todas las posibles funciones que pudieran aplicarse para la solución de dichos problemas pero una clasificación de las mismas ayuda mucho en la consecución de ese fin.

De acuerdo a las características del conjunto imagen L de la función de similaridad, éstas pueden agruparse en:

- booleanas
- k-valentes
- reales

4.1 Funciones booleanas

Las funciones de similaridad booleanas tienen como conjunto imagen $L=\{0,1\}$, que se interpreta como: si el valor es 0, los objetos son diferentes y si el valor es 1, son semejantes.

Veamos algunos ejemplos de funciones de similaridad booleanas:

$$\mathcal{I}(I_{/T}(O_i),I_{/T}(O_j))=\begin{cases}1 & si\ \forall x_p \in T, x_p(O_i)=x_p(O_j)\\ 0 & en\ otro\ caso\end{cases},$$

esta función de similaridad es la denominada *igualdad simple*.

$$\mathcal{I}(I_{/T}(O_i),I_{/T}(O_j))=\begin{cases}1 & si\ \left|\left\{x_p \in T \middle| C_p(x_p(O_i),x_p(O_j))=0\right\}\right| \le \varepsilon\\ 0 & en\ otro\ caso\end{cases},$$

siendo las funciones C_p criterios de comparación de similaridad por variable y ε un umbral.

$$\mathcal{I}(I_{/T}(O_i),I_{/T}(O_j))=\begin{cases}1 & si\ \left|\left\{x_p \in T \middle| C_p(x_p(O_i),x_p(O_j))=0\right\}\right| \le \varepsilon_2\ y\\ & \quad \left|\left\{x_p \in T \middle| C_p(x_p(O_i),x_p(O_j))=1\right\}\right| \ge \varepsilon_1\\ 0 & en\ otro\ caso\end{cases},$$

donde ε_2 y ε_1 son parámetros que regulan la cantidad máxima admisible de rasgos diferentes y la cantidad mínima de rasgos coincidentes, respectivamente.

$$\mathcal{I}(I_{/T}(O_i),I_{/T}(O_j))=\begin{cases}1 & si\ el\ \%\ de\ rasgos\ de\ T\ coincidentes\ es \ge \lambda\%\\ 0 & en\ otro\ caso\end{cases},$$

siendo λ un parámetro de umbral.

4.2 Funciones k-valentes

Las funciones de similaridad k-valentes tienen su imagen en el conjunto totalmente ordenado $L=\{0,1,...,k\text{-}1\}$. Los valores 0 y *k-1* se corresponden con los valores de similaridad mínima y máxima respectivamente, el resto de los valores son gradaciones discretas de la similaridad entre un par de objetos. Por ejemplo:

$$\mathcal{I}(I_{/T}(O_i),I_{/T}(O_j))=\left\lfloor \frac{\sum_{p=1}^{|T|} C_p(x_p(O_i),x_p(O_j))}{|T|} \right\rfloor,$$

donde el símbolo $\lfloor\ \rfloor$ denota parte entera y C_p criterios de comparación de similaridad *k*-valuados por variable. Aquí se asume que todas las variables tienen la misma importancia.

En ocasiones al conjunto L se le da un tratamiento trivalente: se elige un $s \in L$, $s<\lfloor k/2 \rfloor$, los valores menores que *s* se consideran *destacados negativos* o análogos al 0 booleano, los mayores que *k-s*-1 *destacados positivos* o análogos al 1 booleano y los valores entre *s* y *k-s*-1 *no destacados*, que nos darán grados de similaridad y/o diferencia.

4.3 Funciones reales

La imagen de estas funciones pertenece al conjunto $\mathbb{R}$ de los números reales. En ocasiones si L es un conjunto acotado se realiza una proyección sobre el intervalo [0,1]. Muchas de las funciones de similaridad reales se definen a partir de distancias, por ejemplo, del tipo $\Gamma(O_i,O_j)=1 - d_p(O_i,O_j)$, donde

$$d_p(O_i,O_j)=\left\{\sum_{k=1}^{n}\left|x_k(O_i)-x_k(O_j)\right|^p\right\}^{\frac{1}{p}},$$

donde O_i y O_j son dos puntos n-dimensionales en $\mathbb{R}^n$.

Son muy socorridos los casos de p=1, 2 y ∞ correspondientes a la distancia *City-block*, la Euclidiana y la *Chessboard* respectivamente. Otros ejemplos pueden ser:

Sea $f_s(O_h)=\frac{x_s(O_h)-x_s^0}{x_s^1-x_s^0}$, donde x_s^1 y x_s^0 son los valores máximo y mínimo, respectivamente, de la variable x_s.

1) $\Gamma(I_T(O_i),I_T(O_j))=\sum_{x_p\in T}\alpha_p\left(\frac{min\left\{f_p(O_i),f_p(O_j)\right\}}{max\left\{f_p(O_i),f_p(O_j)\right\}}\right)^{\lambda_p}$, $\alpha_p>0$, $\lambda_p\geq 1$.

2) $\Gamma(I_T(O_i),I_T(O_j))=\frac{\sum_{x_p\in T}\left(f_p(O_i)f_p(O_j)\right)}{\left[\sum_{x_p\in T}\left(f_p(O_i)\right)^2\right]^{\frac{1}{2}}\left[\sum_{x_p\in T}\left(f_p(O_j)\right)^2\right]^{\frac{1}{2}}}$.

Como se puede apreciar este es un terreno fértil donde, una vez más, las aplicaciones de estos modelos a los problemas del mundo real crean necesidades de elaborar nuevos tipos de funciones de comparación entre descripciones de objeto, en particular, cuando se sigue una adecuada modelación matemática de los problemas en cuestión [Cheremesina, Ruiz-Shulcloper (1992), Ruiz-Shulcloper, et al. (2013)]. Entre las muchas opciones se encuentran los kernels [Schölkopf, Smola, (2002)]. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press.], que además tienen la ventaja adicional de permitir usar modelos exitosos en la práctica, como las máquinas de vectores de soporte y otros clasificadores basados en kernels [Vapnik, (1998)].

5 La regla del vecino más similar

En el contexto del análisis de datos en general, uno de los procedimientos más conocido es la regla del Vecino Más Cercano (NN-Rule), creado por Fix and Hodges [Fix, Hodges, (1951)]. La esencia de esta regla es que un objeto representado en un espacio métrico es asignado a la clase de su vecino más cercano, el cual se determina sobre la base de una cierta función de distancia.

Una de las ventajas de este procedimiento de decisión es que funciona notablemente bien teniendo en cuenta además, que no requiere de ningún conocimiento explícito de la distribución de los datos, excepto que los mismos deben estar normalizados de igual manera en cada una de las dimensiones del espacio de representación.

Por otro lado se demostró en [Cover, Hart (1967)] que el error de este procedimiento está acotado

asintóticamente por el doble del error de Bayes, que es el error mínimo posible de un clasificador desde el punto de vista estadístico.

Para enfrentar la situación de aplicar esta regla a las descripciones de objetos en términos de datos mezclados e incompletos, además de las opciones de codificar las variables cualitativas y tratar los códigos como números, se desarrollaron algunas funciones supuestamente distancias, denominada por los autores como funciones de distancia heterogéneas. Entre estas podemos mencionar la denominada *Heterogeneous Euclidean-Overlap Metric* (*HEOM*) [Wilson, Martinez (1997)] que es una función análoga a la empleada en los algoritmos IB1, IB2 e IB3 [Aha, et al. (1991), Aha, (1992)] y también usada por [Giraud-Carrier, Martinez (1995)]. La distancia entre dos objetos *O* y *O'* se define como:

$$HOEM(O,O') = \sqrt{\sum_{i=1}^{n} C_i^{\,2}(x_i(O), x_i(O'))} \quad ,$$

donde

$$C_i(x_i(O), x_i(O')) = \begin{cases} 1 & si\ x_i(O)\ ó\ x_i(O')\ son\ valores\ faltantes \\ overlap(x_i(O), x_i(O')) & si\ el\ rasgo\ x_i\ es\ no\ numérico \\ \dfrac{|x_i(O) - x_i(O')|}{max_i - min_i} & si\ el\ rasgo\ x_i\ es\ numérico \end{cases} \quad ,$$

siendo max_i y min_i, respectivamente, los valores máximo y mínimo del rasgo x_i en M_i y la función *overlap*

$$overlap(x_i(O), x_i(O')) = \begin{cases} 0 & si\ \ x_i(O) = x_i(O') \\ 1 & en\ otro\ caso \end{cases} .$$

Debemos observar que esta función C_i considera que cuando uno de los valores que se compara es desconocido su imagen es 1, por lo que en ese caso $HOEM(O,O) \neq 0$. Luego la función *HEOM* no es reflexiva, por lo que no es una distancia.

En esa misma dirección, otra función para comparar subdescripciones mezcladas e incompletas de objetos, para ser utilizada en problemas donde los objetos están divididos en clases, fue definida en [Wilson, Martinez (1997)] *Heterogeneous Value Difference Metric* (*HVDM*) como:

$$HVDM(O,O') = \sqrt{\sum_{i=1}^{n} C_i^{\,2}(x_i(O), x_i(O'))} \ ,$$

donde $C_i(x_i(O),x_i(O'))$ se define como:

$$C_i(x_i(O), x_i(O')) = \begin{cases} 1 & si\ x_i(O)\ o\ x_i(O')\ son\ datos\ faltantes \\ VDM(x_i(O), x_i(O')) & \text{si el rasgo } x_i\ es\ no\ numérico \\ \dfrac{|x_i(O) - x_i(O')|}{4\sigma_i} & \text{si el rasgo } x_i\ es\ numérico \end{cases} \quad ,$$

donde σ_i es la desviación estándar del rasgo x_i en M_i y la función *VDM* se define como:

$$VDM(x_i(O), x_i(O')) = \sqrt{\sum_{c=1}^{C}\left(\frac{N_{i,x_i(O),c}}{N_{i,x_i(O)}} - \frac{N_{i,x_i(O'),c}}{N_{i,x_i(O')}}\right)^2},$$

donde $N_{i,x_i(O)}$ es el número de veces que el rasgo x_i tiene valor $x_i(O)$ en el conjunto de entrenamiento; $N_{i,x_i(O),c}$ es el número de veces que el rasgo x_i tiene valor $x_i(O)$ en los objetos del conjunto de entrenamiento que pertenecen a la clase c. Análogamente a la función *HEOM*, esta función tampoco es reflexiva, por lo que no es una distancia.

Además de esta situación con los datos mezclados e incompletos, como ya hemos ejemplificado anteriormente, existen en la práctica funciones de similaridad que no son ni el inverso ni el opuesto de distancia alguna, o que no son simétricas, o que no son reflexivas, o que no cumplen con la desigualdad triangular, o que los datos son mezclados e incompletos lo que hace imposible la normalización de los mismos o el uso de las llamadas distancias heterogéneas como las mencionadas anteriormente, por lo que para esos casos no tendría sentido aplicar la regla del vecino más cercano.

Con el propósito de emplear un procedimiento análogo al del Vecino Más Cercano, se ha desarrollado una extensión de este procedimiento: la *Regla del Vecino Más Similar* [García-Borroto, Ruiz-Shulcloper (2005)].

Consideremos como antes un universo de objetos U, estructurado en r clases $K_1,\ldots,K_r$ y descritos en términos de un conjunto finito de variables $R=\{x_1,\ldots,x_n\}$ cada una de las cuales tiene asociado un conjunto de valores admisibles M_i, el cual incluye el símbolo '*' para denotar el valor faltante (perdido, desconocido, *missing value*). Sobre los conjuntos M_i no se asume estructura algebraica, topológica o lógica alguna. Por lo que $U=M_1\times\ldots\times M_n$, es sólo el producto cartesiano de los conjuntos de valores admisibles de las variables de R. Sea $O=(x_1(O),\ldots, x_n(O))$, donde $x_i:U\rightarrow M_i$. Cada variable tiene asociado un criterio de comparación de valores $C_i:M_i\times M_i\longrightarrow L_i$ siendo los L_i conjuntos totalmente ordenados. Sea Γ una función de similaridad y $\alpha(O)=(\alpha_1(O),\ldots,\alpha_r(O))$ la tupla de pertenencia de O donde cada $\alpha_i(O)$ denota el grado de pertenencia de O la clase K_i, $i=1,..,r$. Sea $Q=\cup_{i=1}^{r} K_i'$, donde $K_i'\subseteq K_i$, $i=1,\ldots,r$ es un conjunto de entrenamiento. Dado un objeto $O\in U\backslash Q$, la *Regla del Vecino Más Similar* (*MSN* por su sigla en inglés) para clasificar a O consiste en asignarle la tupla de pertenencia $\alpha(O)$ de la manera siguiente:

A) Asumiendo que Γ es solo una función de similaridad

Si $\max\left\{\max_{O_i\in Q}\{\Gamma(O,O_i)\}, \max_{O_i\in Q}\{\Gamma(O_i,O)\}\right\}=\Gamma(O',O)$, entonces $\alpha(O)=\alpha(O')$

B) Asumiendo que Γ es una función de similaridad simétrica

Si $\max_{O_i\in Q}\{\Gamma(O,O_i)\}=\Gamma(O',O)$, entonces $\alpha(O)=\alpha(O')$

siendo $O'\in Q$.

Observe que en estos casos la regla MSN no requiere que el conjunto de las clases forme una partición ni que el universo de objetos tenga una estructuración dura, es decir, en términos de la Teoría Clásica de Conjuntos.

Este procedimiento es fácilmente entendible al caso de los k-vecinos más similares. La regla de los k *Vecinos Más Similares* (*k-MSN* por su sigla en inglés) se basa en construir la tupla de pertenencia del objeto a clasificar sobre la base de las respectivas tuplas de pertenencia de los k objetos más similares a dicho objeto. Aquí vamos a diferenciar 4 posibles situaciones que dependen del tipo de estructuración que presente el conjunto de clases del problema en cuestión: particiones duras, cubrimientos duros, cubrimientos difusos y particiones difusas de Ruspini.

Sea dado un objeto $O\in U\backslash Q$, y sean $\alpha(O_i)=(\alpha_1(O_i),\ldots,\alpha_r(O_i))$, $i=1,\ldots,k$, las tuplas de pertenencia de los respectivos k vecinos más cercanos a O en Q, donde $\alpha_j(O_i)$ representa el grado de pertenencia del objeto O_i a la clase K_j, $j=1,\ldots,r$.

La regla de los k-MSN para asignar al objeto O su tupla de pertenencia $\alpha(O)$ es la siguiente:

A) Si el conjunto de clases $K=\{K_1,..,K_r\}$ forma una partición dura, $|\alpha(O_i)|=1$, $\alpha_j(O_i)\epsilon\{0,1\}$, $j=1,\ldots,r$, $i=1,\ldots,k$, es decir, cada objeto pertenece sólo a una clase; la regla k-MSN en este caso le asigna a O la clase más representada en las tuplas de pertenencia de los k-vecinos más cercanos:

Si $\max_{j=1,\ldots,r}\{\sum_{i=1}^{k} \propto_j (O_i)\} = \sum_{i=1}^{k} \propto_h (O_i)$, entonces

$$\propto_j (O) = \begin{cases} 1 & para\ i = h \\ 0 & en\ otro\ caso \end{cases}, j=1,\ldots,r.$$

B) Si el conjunto de clases $K=\{K_1,..,K_r\}$ forma un cubrimiento duro, $|\alpha(O_i)|>1$, $\alpha_j(O_i)\epsilon\{0,1\}$, $j=1,\ldots,r$, $i=1,\ldots,k$, es decir, cada objeto puede pertenecer simultáneamente a más de una clase, la decisión en este caso se toma independientemente para cada clase, asignándole a O la clase K_j si en el conjunto de las tuplas de pertenencia de los k-vecinos más cercanos al menos Δ veces aparece la clase K_j:

$$\alpha_j(O) = \begin{cases} 1 & \sum_{i=1}^{k} \propto_j (O_i) \geq \Delta \\ 0 & en\ otro\ caso \end{cases},$$

donde $\Delta\epsilon\{0,1,\ldots,k\}$, para j=1,…,r.

C) Si el conjunto de clases $K=\{K_1,..,K_r\}$ forma un cubrimiento difuso, $|\alpha(O_i)|>1$, $\alpha_j(O_i)\epsilon[0,1]$, $j=1,\ldots,r$, $i=1,\ldots,k$, es decir, cada objeto pertenece a cada clase en un cierto grado, la regla en este caso asigna a O la media de los grados de pertenencia de sus k-vecinos más cercanos a cada una de las clases:

$$\propto_j (O) = \frac{\sum_{i=1}^{k} \propto_j (O_i)}{k}.$$

D) Si el conjunto de clases $K=\{K_1,..,K_r\}$ forma una partición difusa de Ruspini, $|\alpha(O_i)|=1$, $\alpha_j(O_i)\epsilon[0,1]$, $j=1,\ldots,r$, $i=1,\ldots,k$, es decir, cada objeto pertenece a cada clase en un cierto grado pero la suma de esos grados de pertenencia tiene que ser igual a 1. En este caso la regla procede como en el caso del cubrimiento difuso, pero normalizando la tupla de pertenencia:

$$\propto_j (O) = \frac{\sum_{i=1}^{k} \propto_j (O_i)}{\sum_{k=1}^{r} \sum_{i=1}^{k} \propto_k (O_i)}.$$

Por ejemplo, supongamos que el universo en cuestión lo constituye un conjunto de noticias, con tres clases difusas: las noticias de guerra, las económicas y las políticas. Consideremos tres noticias en particular: O_1, O_2 y O_3 que son las noticias más similares a la noticia O que queremos clasificar, sus tres vecinos más similares. Asumamos que las tuplas de pertenencia respectivas son $\alpha(O_1)$=(0.8, 0.7, 0.1); $\alpha(O_2)$=(0.7, 0.5, 0.3); $\alpha(O_3)$=(0.2, 0.6, 0.4). Entonces, en caso que asumamos que el conjunto de noticias en estudio forman un cubrimiento difuso, la regla MSN asignaría a la noticia O, la tupla de pertenencia $\alpha(O)$=(0.57, 0.6, 0.27) queriendo expresar que es una noticia de guerra con fuertes implicaciones económicas pero con una relevancia menor desde el punto de vista político.

Por su parte si el caso que nos ocupara fuese el de una partición difusa de Ruspini, entonces la tupla de pertenencia sería $\alpha(O)$=(0.39, 0.42, 0.19).

Como ya habíamos subrayado, la regla del Vecino Más Cercano o su extensión a los k-Vecinos Más Cercanos, es una herramienta con muchos atractivos para su utilización en la práctica, que ha sido aplicada en una extensa gama de problemas reales, en ocasiones de manera incorrecta, según nuestro criterio, pero con la única limitación de que asume una función de distancia con todo lo que eso

presupone. La extensión de este procedimiento a la Regla del Vecino Más Similar, y a los *k*-Vecinos Más Similares, da la posibilidad de aprovechar la mayoría de las ventajas antes mencionadas pero además amplía el marco de sus posibles aplicaciones a problemas del mundo real con una fundamentada argumentación de su empleo en condiciones como las que mencionábamos de problemas en los que la analogía debe ser modelada por una función que no es el inverso ni el opuesto de una distancia, que puede no ser simétrica, permitiendo además poder modelar problemas con descripciones de objetos en términos de datos mezclados e incompletos, estos últimos muy frecuentes en disciplinas como la Medicina, las Geociencias, la Criminalística entre otras, en las cuales el procedimiento original no podría ser aplicado.

Con esta extensión del procedimiento *k*-NN, con cualquier función de similaridad que se esté trabajando es factible el uso del Vecino Más Similar y sus extensiones.

En esta dirección se ha continuado trabajando y se han obtenido una serie de resultados que conforman un adecuado arsenal para enfrentar esta problemática [Olvera-López, et al. (2005a); Olvera-López, et al. (2005b); Hernández-Rodríguez, et al. (2007); García-Borroto, et al. (2009)]

6 C-Means con funciones de similaridad

Otro de los procedimientos muy extendido por su uso en Reconocimiento de Patrones, en particular en la solución de problemas de clasificación no supervisada restringida, es el C-Means [Ball, Hall, (1967)], debido a su simplicidad, baja complejidad y los buenos resultado que ha reportado en la práctica [Jose et al. (2014), Polczynski & Polczynski (2014)]. Desde su primera publicación este algoritmo ha sido objeto de múltiples extensiones, sin embargo, esos nuevos desarrollos han sido fundamentalmente atendiendo a la selección de las semillas iniciales [Bradley, Fayyad (1998).], la determinación del número óptimo de agrupamientos a formar [Dubes (1987)] y el uso de diferentes funcionales para la generación de los agrupamientos [Bobrowsky, Bezdek, (1991)]. Problemas, como los antes mencionados, con datos mezclados e incompletos o con funciones de similaridad que no son el inverso ni el opuesto de una distancia no han recibido la misma atención.

Como se conoce este algoritmo parte de una partición inicial y posteriormente intercambia los datos de un agrupamiento a otro de manera iterativa con el fin de optimizar una cierta función objetivo. El método presupone que los objetos están descritos en términos de variables a las que se les puede aplicar una función distancia.

En el contexto de los problemas de clasificación no supervisada restringida se han desarrollado algunos algoritmos para enfrentar la situación de los datos mezclados e incompletos y uno de los más representativos es el C-Means conceptual desarrollado por Ralambondrainy en 1995 [Ralambondrainy, (1995)]. En este algoritmo se introduce una distancia para manejar los datos mezclados, sobre la base de calcular la distancia aportada por las variables numéricas, empleando la distancia Euclidiana, más la distancia aportada por las variables cualitativas, empleando para ello la distancia chi-cuadrado, para lo cual cada valor de las variables cualitativas es codificado como una variable binaria. La distancia así introducida es interpretada en el espacio n-dimensional original y se calculan los centroides. Sin embargo, esto es erróneo pues las distancias parciales han sido calculadas en espacios diferentes. Observe que ninguna de las variables numéricas fue evaluada de conjunto con alguna cualitativa para valorar la similaridad entre los objetos. Este hecho en muchas aplicaciones a problemas reales resulta inadmisible.

A esto se une el hecho que los especialistas de las áreas de las ciencias poco formalizadas, donde aparecen con frecuencia los datos mezclados e incompletos y funciones de similaridad con las características antes mencionadas, también tienen que enfrentar problemas de clasificación no supervisada restringida, por lo que se impone lograr una adecuada extensión de métodos como el C-Means [García Serrano, Martínez-Trinidad (1999)].

Consideremos un conjunto de objetos $\{O_1,O_2,...,O_m\}$ que deben ser estructurados en c agrupamientos. Asumimos, como antes, los objetos descritos en términos de un conjunto de n rasgos

(de cualquier naturaleza), cada uno de los cuales tiene asociado un conjunto de valores admisibles M_i y un criterio de comparación de valores C_i, i=1,…,n . Sea dada una función de similaridad Γ esta función pudiera depender de cierta manera de funciones de similaridades parciales, es decir, de funciones $\Gamma' : (M_{i_1} \times ... \times M_{i_s})^2 \rightarrow L'_i$, con L'_i totalmente ordenados y s<n. Estas funciones de similaridad parcial nos permiten comparar subdescripciones de objetos que están dadas en términos de un subconjunto de variables. Este subconjunto se denomina *conjunto de apoyo* y el conjunto de conjuntos de apoyo *sistema de conjuntos de apoyo*. Asumamos que Γ es una función de similaridad simétrica que toma valores en el conjunto [0,1]. Denotemos $\alpha_i(O_k)$ el grado de pertenencia del objeto O_k al agrupamiento C_i, y consideremos el conjunto R_{cxm} de todas las matrices reales de cxm. Así, cualquier partición de $\{O_1,O_2,...,O_m\}$ puede ser representada por una matriz $[\alpha_i(O_k)] \in R_{cxm}$ que satisface las siguientes condiciones:

1. $\alpha_i(O_k) \in \{0,1\}$ k=1,…,m; i=1,…,c.
2. $\sum_{i=1}^{c} \alpha_i(O_k) = 1$ k=1,..,m.
3. $\sum_{k=1}^{m} \alpha_i(O_k) > 0$ i=1,…,c.

La matriz de partición H=$[\alpha_i(O_k)]$ será determinada a partir de maximizar la función objetivo $J(H)=\sum_{i=1}^{c}\sum_{k=1}^{m} \alpha_i(O_k)\Gamma(O_i^r, O_k)$, siendo O_i^r el objeto más representativo (el prototipo, el holotipo, el "centro") en el agrupamiento C_i. Para determinar este objeto más representativo en cada agrupamiento, que lo denominaremos en lo sucesivo el *holotipo* del agrupamiento, dada la matriz de partición H consideraremos la siguiente función

$$r_{C_i}(O_j) = \frac{\beta_{C_i}(O_j)}{\left(\alpha_{C_i}(O_j) + \left(1 - \beta_{C_i}(O_j)\right)\right)} + \eta_{C_q}(O_j), \qquad (1)$$

donde:

$O_j \in C_i$, $C_q \neq C_i$,

$$\beta_{C_i}(O_j) = \frac{1}{|C_i| - 1} \sum_{\substack{O_j, O_q \in C_i \\ O_j \neq O_q}} \Gamma(O_j, O_q), \qquad (2)$$

$$\alpha_{C_i}(O_j) = \frac{1}{|C_i| - 1} \sum_{\substack{O_j, O_q \in C_i \\ O_j \neq O_q}} \left| \beta_{C_i}(O_j) - \Gamma(O_j, O_q) \right|, \qquad (3)$$

$$\eta_{C_k}(O_j) = \sum_{\substack{q=1 \\ i \neq q}}^{c} \left(1 - \Gamma(O_q^r, O_j)\right). \qquad (4)$$

La función $\beta_{C_i}(O_j)$ evalúa el promedio de la similaridad del objeto O_j con los restantes objetos del agrupamiento C_i. La ecuación (3) nos permite evaluar la varianza entre la media calculada por (2) y la similaridad de O_j con respecto a los otros objetos de C_i. De esta manera, cuando (3) decrece el valor de (1) se incrementa. La expresión $\left(1 - \beta_{C_i}(O_j)\right)$ representa el promedio de disimilaridad del objeto O_j con los restantes objetos de C_i y la expresión (4) evalúa la dismilaridad entre el objeto O_j y los holotipos de los restantes agrupamientos. El objetivo de esta función es disminuir los casos donde existan dos objetos con el mismo valor de (1).

Cuando $|C_i|$=1, el holotipo para el agrupamiento C_i es el objeto contenido en ese agrupamiento.

Por todo lo anterior, es razonable que el holotipo de un agrupamiento se defina como el objeto O_r en el cual se alcanza el máximo valor de $r_{C_i}(O_j)$, esto es

$$r_{C_i}(O_r) = \max_{O_p \in C_i} \{r_{C_i}(O_p)\}. \quad (5)$$

Dado los holotipos de los agrupamientos, el funcional; *J(H)* se maximiza cuando

$$\alpha_i(O_k) = \begin{cases} 1 & if\ \Gamma(O_i^r, O_k) = \max_{1 \le q \le c} \{\Gamma(O_q^r, O_k)\} \\ 0 & en\ otro\ caso \end{cases}. \quad (6)$$

Esto significa que un objeto O_k será asignado al agrupamiento cuyo holotipo es el más similar a O_k.

Sobre la base de estas consideraciones los autores de [García-Serrano, Martínez-Trinidad (1999)] desarrollaron el algoritmo:

C-Means para funciones de similaridad (**SF C-means**)

Paso 1. Sea dado *c*, $2 \le c \le n$. Sea dado un número de iteraciones *ni'*, y *ni*=0.

Paso 2. Seleccionar *c* objetos del conjunto inicial como semillas.

Paso 3. Calcular la matriz de partición $H=H^{(ni)}$ usando (6).

Paso 4. Determinar los holotipos de los agrupamientos para la matriz $H^{(ni)}$, usando (1) y (5).

Paso 5. Si el conjunto de los holotipos es el mismo que en la iteración anterior, detener el proceso, en otro caso incrementar *ni=ni+1*.

Paso 6. Si *ni>ni'* detener el proceso. En caso contrario, ir al Paso 3.

Hasta aquí asumimos que la función de similaridad era simétrica. Dado el caso que la función no cumpla esta propiedad, en las expresiones (2), (3), (4) y (6) se sustituye $\Gamma(O_j, O_q)$ por $\max\{\Gamma(O_q, O_j), \Gamma(O_j, O_q)\}$ y se prosigue como en el caso de simetría.

Como apuntan sus autores, el algoritmo SF C-Means tiene diferencias importantes respecto al algoritmo clásico y da posibilidades de enfrentar situaciones en las que un procedimiento como el C-Means es deseable, pero que sin una extensión como esta sería imposible su aplicación, al menos con sentido.

Esta nueva familia de algoritmos para la solución de problemas de clasificación no supervisada restringida permite: trabajar en espacios de representación de los objetos no métricos, e incluso en simples productos cartesianos, permite el uso de criterios de comparación de valores de cada una de las variables en términos de las cuales se describen a los objetos del universo en cuestión, no necesariamente el mismo para todas las variables y también funciones de similaridad que no sean el inverso o el opuesto de una distancia, y como se aprecia en este trabajo, incluso para funciones de similaridad que no sean simétricas, es decir, simples funciones de similaridad. Además de la posibilidad de comparar subdescripciones de los objetos sobre la base de conjuntos soporte. Por lo antes mencionado, el SF-C-Means puede ser aplicado a datos mezclados e incompletos

Otros trabajos de interés en esta misma dirección pueden ser encontrados en [Martínez-Trinidad, et al. (2002); López-Escobar, et al. (2005); López-Escobar, et al. (2006)].

En este algoritmo SF-C-Means, un elemento que tiene un valor propio, una utilidad en sí, es el *procedimiento para el cálculo de los holotipos de un conjunto de objetos*. Es muy común, no sólo por problemas de optimización de los procedimientos que se aplican a un conjunto de objetos, la necesidad de determinar un representante de dicho conjunto: la foto más representativa, el modo operandi de un conjunto de hechos, la conducta promedio, habitual, de un grupo de personas, el perfil de grupos sociales, y muchos más. En muchos de esos problemas la solución puede buscarse en un espacio métrico, pues los objetos pueden ser definidos en tales espacios, pero muchos otros problemas no son así y se requiere de procedimientos como los descritos en este epígrafe, en particular el de la determinación del holotipo.

7 CSE: un algoritmo para la selección de prototipos

Enfrentando este problema de la selección de prototipos en espacios de representación que no son más que simples productos cartesianos, en [García-Borroto, Ruiz-Shulcloper, (2005)] se introduce un método CSE (por su sigla en inglés: Compact Set Editing) basado en la extensión de la Regla del Vecino Más Cercano para datos mezclados e incompletos y funciones de similaridad que no son el inverso o el opuesto de distancias que obtiene prototipos con características de interés que veremos a continuación. Aunque el CSE fue elaborado con el objetivo de reducir el tamaño de la matriz de entrenamiento en un problema de clasificación supervisada obteniendo una eficacia del clasificador nunca inferior a la obtenida con la matriz de entrenamiento inicial (problema de la edición de matrices de entrenamiento en problemas de clasificación supervisada), sin dudas da solución también al problema de la selección de prototipos no sólo con este fin, sino también como un objetivo en sí, atendiendo a ejemplos como los mencionados en el párrafo anterior. Por otro lado este método permite trabajar también con datos mezclados e incompletos (MID) y funciones menos exigentes que las distancias.

Consideremos como antes una función de similaridad Γ, diremos que O_i, $O_j \in U$ son β_0-*similares* si se cumple que $\Gamma(O_i,O_j) \geq \beta_0$, siendo β_0 un valor de umbral que permite controlar cuán similares deben ser dos objetos para ser considerados β_0-similares. Análogamente diremos que O_i es un objeto β_0-*aislado* si $\forall O_j \neq O_i \in U \;\; \Gamma(O_j,O_i) < \beta_0$.

Siguiendo [Martínez-Trinidad, et al. (2000)] diremos que $NU \subseteq U$, $NU \neq \varnothing$ es un *conjunto* β_0-*compacto* si se cumple:

- $\forall O_j \in U \left[O_i \in NU \wedge \max\limits_{\substack{O_i \in U \\ O_t \neq O_i}} \{\Gamma(O_i,O_t)\} = \Gamma(O_i,O_j) \geq \beta_0 \right] \Rightarrow O_j \in NU$.
- $\left[\max\limits_{\substack{O_i \in U \\ O_i \neq O_{pi}}} \{\Gamma(O_p,O_i)\} = \Gamma(O_p,O_t) \geq \beta_0 \wedge O_t \in NU \right] \Rightarrow O_p \in NU$.
- siendo $|NU|$ minimal.

Además, diremos que todo objeto β_0-aislado forma un conjunto compacto (degenerado).

Basado en este concepto se genera un criterio de agrupamiento que forma una única partición del universo en cuestión. Esa partición tiene la propiedad que un objeto O y todos los que son sus más β_0-similares vecinos, están en el mismo agrupamiento y además están también todos aquellos objetos que tienen a dicho objeto como el vecino más β_0-similar.

Algoritmo CSE

Aquí no consideraremos los detalles relativos al problema de edición y sólo atenderemos a lo concerniente con la selección de los prototipos.

Entrada: Los conjuntos β_0-compactos dados cada uno por un grafo de máxima similaridad, que es una grafo orientado donde cada arista de un vértice *a* hacia un vértice *b* significa que este último es el elemento más β_0-similar al vértice *a*, siendo *C* el conjunto de aristas y *V* el conjunto de vértices de dicho grafo orientado.

Salida: Subconjunto *R* de los prototipos seleccionados

Denotaremos $S(x)=\{b \in V | \; (x,b) \in C\}$, al conjunto de los sucesores del vértice *x* en el grafo y por $A(x)=\{a \in V | \; (a,x) \in C\}$ al conjunto de los antecesores del vértice *x* en el grafo.

Sea $R=\varnothing$

Paso 1. Cada vértice $x \in V$ se asocia con un cuádruplo $(S'_x, E_x, S_x, Flags_x)$, donde

S'_x=|{$y \in S(x)$| $\alpha(x) \neq \alpha(y)$}|
E_x=|{$y \in A(x)$| $\alpha(x)=\alpha(y)$}|
S_x=|{$y \in S(x)$| $\alpha(x)=\alpha(y)$}|
$Flags_x \subset V$, $Flags_x=\varnothing$

Paso 2. $R'=\{x \in V | S'_x>0\}$
Paso 3. Si $R'=\varnothing$ ir al Paso 6
Paso 4. $C \leftarrow C \setminus \{(x,y) \in C | x \in R' \wedge \alpha(x) \neq \alpha(y)\}$
Paso 5. Para cada elemento $x \in R'$ ejecutar *Move*(x)
Paso 6. $\forall x \in V[(S_x=0) \Rightarrow$ejecutar *Move*($x$)]
Paso 7. $\forall x \in V \forall y \in Flags_x \left[y \notin \bigcup_{z \in V \setminus \{x\}} Flags_z \Longrightarrow ejecutar\ Move(x) \right]$
Paso 8. Ordenar los elementos de V con la siguiente relación de orden:

$$x \prec y \Leftrightarrow E_x < E_y \vee \left(E_x = Ey \wedge S_x < S_y\right) \vee \left(E_x = Ey \wedge S_x = S_y \wedge |Flags_x| > \left|Flags_y\right|\right)$$

Paso 9. Ejecutar *Discard*(x_1), donde x_1 es el primer vértice de V respecto al orden definido.
Paso 10. Si $V=\varnothing$ terminar, de lo contrario ir al Paso 6.
Siendo *Move*(x) y *Discard*(x):

***Move*(x)**

M1. Calcular $A(x)$ y $S(x)$ con el conjunto actual de vértices V.
M2. $\forall y \in A(x)[S_y \leftarrow \infty]$
M3. $\forall y \in S(x)[E_y \leftarrow E_y - 1]$
M4. $\forall y \in V[Flags_y \leftarrow Flags_y \setminus Flags_x]$
M5. $V \leftarrow V - \{x\}$, $R \leftarrow R \cup \{x\}$
M6. $C \leftarrow C \setminus \{(a,b) \in C | a=x \vee b=x\}$

***Discard*(x)**

D1. Calcular $A(x)$ y $S(x)$ con el conjunto actual de vértices V.
D2. $\forall y \in S(x)[Flags_y \leftarrow Flags_y \cup \{x\}]$
D3. $\forall y \in S(x)[Ey \leftarrow Ey - 1]$
D4. $\forall y \in A(x)[Sy \neq \infty \Rightarrow Sy \leftarrow Sy\text{-}1$
D5. $V \leftarrow V - \{x\}$
D6. $C \leftarrow C \setminus \{(a,b) \in C | a=x \vee b=x\}$

Como puede observarse, los índices que se calculan en el primer paso del algoritmo constituyen la base fundamental de las decisiones en cuanto a cuáles vértices descartar y cuáles seleccionar. En los pasos del 2 al 6 se conforman los conjuntos compactos por clases, recordemos que este algoritmo fue diseñado para editar las matrices de entrenamiento en problemas de clasificación supervisada. En los restantes pasos se toman las decisiones relativas a los vértices que deben ser removidos del conjunto restando los prototipos correspondientes a cada uno de los conjuntos compactos en cada una de las clases del problema en cuestión.

8 Conclusiones

Como se ha podido apreciar, un aspecto esencial en el proceso de reconocer patrones, en todas sus variantes: diagnóstico, pronóstico, génesis, identificación, selección de variables, etc., es el modo en que los datos son procesados. ¿Cómo se representan los objetos en estudio? y ¿Cómo se comparan dichas descripciones?, son elementos cruciales en la calidad de los resultados finales.

Durante toda la etapa inicial de su desarrollo, el Reconocimiento de Patrones se apoyó, y se sigue apoyando fuertemente, en los modelos creados sobre espacios reales, vectoriales, métricos, es decir, sobre espacios matemáticos con estructuras bien estudiadas, sobre los cuales se ha desarrollado un

importante arsenal de herramientas que permiten un adecuado procesamiento de los datos. Sin embargo, la realidad es mucho más rica y cambiante. La necesidad de resolver problemas que se apartan de esos tradicionales estudios ha obligado a la comunidad científica a plantearse la necesidad de seguir ampliando las posibilidades de abordar estos problemas ahora en condiciones diferentes, para las cuales es necesario crear nuevos recursos matemáticos y computacionales.

Los conceptos de datos mezclados e incompletos y de función de similaridad, no necesariamente una función inversa u opuesta a una distancia, ni siquiera necesariamente una función simétrica, se han utilizado para resolver problemas de Reconocimiento de Patrones y Minería de Datos bajo estas nuevas circunstancias. El surgimiento del Reconocimiento Lógico Combinatorio de Patrones y la Minería de Datos Mezclados e Incompletos [Martínez-Trinidad, Guzmán-Arenas, (2001), Ruiz-Shulcloper, Abidi (2002), Ruiz-Shulcloper, J. (2008)] en el marco respectivo de esas disciplinas, ha abierto un interesante escenario de aplicaciones y de investigaciones teóricas que buscan crear modelos, conceptos y herramientas necesarias para enfrentar con eficacia estos tradicionales problemas pero bajo nuevas condiciones. Es por ello un área fértil para el desarrollo de investigaciones teóricas y de aplicaciones de estas nuevas herramientas a la solución de problemas en importantes áreas del conocimiento que por demás tienen un impacto social y económico nada despreciable.

Referencias bibliográficas

1. Aha, D. W., Kibler, D., Albert, M. K. (1991). Instance-based learning algorithms, Machine Learning, vol. 6.
2. Aha, D. W., (1992). Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. International Journal of Man-Machine Studies, Vol. 36, pp. 267-287.
3. Alba-Cabrera, E. (1998). New extensions of the testor concept for different types of similarity functions. PhD dissertation, Instituto de Cibernética, Matemática y Física, Cuba.
4. Álvarez-Gómez, L., Ruiz-Shulcloper, J., Chuy-Rodríguez, T., Pico-Peña, R., Cotilla, M. (1992). Modelación Matemática del Pronóstico de Magnitudes Máximas de los Terremotos en la Región del Caribe. En: Editores J. Ruiz-Shulcloper, L. Álvarez-Gómez, V. Guitis. Reconocimiento de Estructuras Espaciales. pp 81-101, Editorial Academia. Cuba.
5. Ayaquica-Martínez, I. O., Martínez-Trinidad, J. F. Carrasco-Ochoa, J. A. (2006). Conceptual K-Means Algorithm based on Complex Features, Lecture Notes in Computer Science 4225, Springer, 11th Iberoamerican Congress on Pattern Recognition (CIARP2006), Cancún, México, Noviembre, 2006, pp. 491-501.
6. Ball, G., Hall, D. A Clustering technique for summarizing multivariate data. Behav. Sci. (12) 153-155 (1967).
7. Baskakova, L.V. Zhuravlev, Yu.I. (1981). Model of algorithm of recognition with representative sets and support sets systems, Zhurnal Vichislitielnoi Matemati y Matematicheskoi Fisiki, vol. 21, no.5. pp. 1264-1275.
8. Bobrowsky, L., Bezdek, J.C. (1991). C-means clustering with the L1 and L∞ Norms. IEEE Transactions on Systems man, and Cybernetics, 21, 3, pp. 545-554.
9. Bongard, M.N. (1963). Solution to geological problems with support of recognition programs, Sov. Geologia, vol. 6, pp. 33-50.
10. Bradley, P., Fayyad, U. (1998). Refining Initial Points for K-Means Clustering. In Proceedings of the Fifteenth International Conference on Machine Learning ICML98, Morgan Kaufmann, San Francisco, pp. 91-99.
11. Chen, H., Lynch, K.J. (1992). Automatic construction of networks of concepts characterizing document databases. IEEE Transaction on Systems, Man, and Cybernetics, vol. 22, 5, September/October, pp. 885-902.
12. Chen, H., Ng, T.D., Martínez, J., Schatz, B.R. (1997). A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the worm community system. Journal of the American Society for Information Science, 48(1), pp. 17-31.
13. Cheremesina, E.N., Ruiz-Shulcloper, J. (1992). Cuestiones Metodológicas de la Aplicación de Modelos Matemáticos de Reconocimiento de Patrones en Zonas del Conocimiento Poco Formalizadas. Revista Ciencias Matemáticas, vol 13; No.2; pp. 93-108, Cuba.

14. Chidananda-Gowda, K., Ravi, T. V. (1995). Agglomerative clustering of symbolic objects using the concepts of both similarity and dissimilarity", Pattern Recognition Letters 16, pp. 647-652.
15. Cover, T. M., Hart, P. E. (1967). Nearest Neighbor Pattern Classification, IEEE Transactions on Information Theory, vol. 13, pp. 21-26.
16. Douglas-De la Peña, M., Ruiz-Shulcloper, J. (1983). Un Algoritmo para el Pronóstico de Enfermedades Laborales Crónicas. Revista Ciencias Matemáticas Vol. IV(1) pp.133-155. Cuba.
17. Dubes R. C. (1987). How many clusters are best? -An experiment Pattern Recognition, 20, pp. 645-663.
18. Fix, E., Hodges, J. (1951). Discriminatory Analysis. Nonparametric discrimination: Consistency properties. Technical Report 4, USAF School of Aviation Medicine.
19. García-Borroto, M., Ruiz-Shulcloper, J. (2005). Selecting Prototypes in Mixed Incomplete Data. Lecture Notes in Computer Science, Lecture Notes in Computer Science 3773, pp. 450–459.
20. García-Borroto, M., Medina-Pérez, M.A., Villuendas-Rey, Y., Ruiz-Shulcloper, J. (2009). Búsqueda rápida del vecino más similar en espacios no métricos. Revista Cubana de Ciencias Informáticas, Vol.3, No.1-2, pp. 5-11.
21. García-Serrano, J. R., Martínez-Trinidad, J.F. (1999). Extension to c-means algorithm for the use of similarity functions. 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases Proceedings. Prague, Czech Rep. pp. 354-359.
22. Giraud-Carrier, Ch., Martinez, T. R. (1995). An Efficient Metric for Heterogeneous Inductive Learning Applications in the Attribute-Value Language. Intelligent Systems, pp. 341-350.
23. Goldfarb, L. (1985). A new approach to Pattern Recognition. In: Progress in Machine Intelligence & Pattern Recognition. Ed. L. Kanal; A. Rosenfeld, Vol II., pp. 241-402.
24. Gómez-Herrera, J.E., Rodríguez-Morán, O., Valladares-Amaro, S., Ruiz-Shulcloper, J., Pico-Peña, R., Echevarría-Rodríguez, G., Tenreyro-Pérez, R., Otero-Marrero, R., Cheremisina, E.N., Cruz-Toledo, R., Barceló-Carol, G., Álvarez-Castro, J., Barea-Centeno, M., García-Sánchez, R. (1994). Pronóstico Gasopetrolífero en la Asociación Ofeolítica Aplicando la Modelación Matemática. Revista Geofísica Internacional, Volumen 33, No. 3, July-Sept., pp 447-467. México
25. Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, 53, No. 3, 4, pp. 325-338.
26. Gower, J.C. (1967). A comparison of some methods of cluster analysis. Biometrics, December, pp. 623-637.
27. Gower, J.C. (1971). A general coefficient of similarity and some its properties. Biometrics, 27, December, pp. 857-871.
28. Gower, J.C. (1977). The analysis of asymmetriy and orthogonality, in Recent Developments in Statistics, (J.R. Barra, F. Brodeau, G. Romier and B. van Cutsem, eds.), North Holland Publishing Company, Amsterdam, pp. 109-123.
29. Hernández-Rodríguez S., Carrasco-Ochoa, J. A., Martínez-Trinidad, J. F. (2008). Fast k Most Similar Neighbor Classifier for Mixed Data based on a Tree Structure and Approximating-Eliminating, Lecture Notes in Computer Science 5197, Springer. 13th Iberoamerican Congress on Pattern Recognition (CIARP2008), Habana, Cuba, Septiembre, 2008, pp. 364-371.
30. Hernández-Rodríguez, S., et al. (2007). Fast Most Similar Neighbor Classifier for Mixed Data, Lecture Notes in Computer Science 4509, pp. 146-158
31. Hubert, L.J. (1973). Min and max hierarchical clustering using asymmetric similarity measures. Psychometrika, No. 38, pp. 63-72.
32. Jain, K., Dubes, R.C. (1998). Algorithms for Clustering Data, Prentice Hall.
33. Jose, A., Ravi, S., & Sambath, M. (2014). Brain Tumor Segmentation Using K-Means Clustering And Fuzzy C-Means Algorithms And Its Area Calculation. Brain, 2(3).
34. Kochetkov, D.V. (1978). About nearness function. Apply Mathematics Communications, Computer Center, Technical Report of Academy of Sciences SSSR, pp. 1-30.
35. Kodratov, Y., Tecuci, G. (1988). Learning based on conceptual distance. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 10, No. 6, pp. 897-909.
36. López-Reyes, N., Ruiz-Shulcloper, J., Gil-Moreno, G., Viera, L. (1988). Un Sistema para el Pronóstico a Corto Plazo de Tormentas Ionosféricas. Reporte de Investigación ICIMAF, No.76, pp 1-25. Cuba.
37. López-Escobar, S., Carrasco-Ochoa, J. A., Martínez-Trinidad, J.F. (2005). Global k-Means with Similarity Functions. Progress in Pattern Recognition, Image Analysis and Applications, Lecture Notes in Computer Science 3773, pp. 392-399.
38. López-Escobar, S., Carrasco-Ochoa, J. A., Martínez-Trinidad, J.F. (2006). Fast Global k-Means with Similarity Functions Algorithm. Lecture Notes in Computer Science 4224, pp. 512–521.

39. Mahalanobis, P.C. (1936). On the generalized distance in statistics, Proceedings of the National Institute of Science of India 12, pp. 49-55.
40. Martínez-Trinidad, J.F., Ruiz-Shulcloper, J., Lazo-Cortés, M. (2000). Structuralization of Universes. Fuzzy Sets and Systems, Vol. 112, No. 3, pp. 485-500.
41. Martínez-Trinidad, J.F., García-Serrano, J. R., Ayaquica-Martínez, I. O. (2002). C-Means Algorithm with Similarity Functions, Computación y Sistemas Vol. 5 No. 4, pp. 241-246
42. Michalski, R.S. (1969). Algorithm Aq for the Quasi-Minimal Solution of the General Covering Problem. Journal Archiwum Automatyki i Telemechaniki, 4, Polish Academy of Sciences.
43. Moskalienskii, E.D. (1984). On the construction of similarity measure that not fulfill the redundant symmetric property, in Computational Methods of Geology, Novosibirsk, pp. 90-105.
44. Moskalienskii, E.D., Chinaiev, Y.B. (1984). On classification similarity functions between objects described in Boolean variables, in Computational Methods of Geology, Novosibirsk, pp. 157-161.
45. Olvera-López, J. A., Carrasco-Ochoa, J. A., Martínez-Trinidad, J. F. (2005a). Sequential search for decremental edition. Lecture Notes in Computer Science, vol. 3578, pp. 280-285.
46. Olvera-López, J. A., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A. (2005b). Edition schemes based on BSE. Lecture Notes in Computer Science, vol. 3773, pp. 360-367.
47. Ortíz-Posadas, M.R. (1997a). Prognosis and evaluation of cleft palate patients' rehabilitation using pattern recognition techniques. Proceedings of World Congress on Medical Physics and Biomedical Engineering 35, 1, pp. 500-. Niza, France.
48. Ortíz-Posadas, M.R., Lazo-Cortés, M. (1997b). Evaluation of cleft palate patients' rehabilitation using pattern recognition techniques. Proceedings of II Taller Iberoamericano de Reconocimiento de Patrones. Conferencia Internacional CIMAF'97. La Habana, pp. 231-236.
49. Ortíz-Posadas, M.R., Martínez-Trinidad, J.F., Ruiz-Shulcloper, J. (1996). A new approach to differential diagnosis of diseases. International Journal of Biomedical Computing, 40, 3, pp. 179-185.
50. Ortíz-Posadas, M.R., Maya-Behart, J., Lazo-Cortés, M. (1998a). Evaluation of lips and cleft palate chirurgery using logical combinatorial approach to pattern recognition theory. Journal Revista Brasileira de Bioengenharia. Caderno de Engenharia Biomedica, 14, No. 1, pp. 7-21.
51. Ortíz-Posadas, M.R., Vega-Alvarado, L., Jiménez-Jacinto, V., Lazo-Cortés, M. (1998b). The concept of analogy in medicine. A similarity function for cleft palate patients. Proceedings of III Taller Iberoamericano de Reconocimiento de Patrones. México, pp 247-256.
52. Ortíz-Posadas, M.R., Vega-Alvarado, L., Jiménez-Jacinto, V., Lazo-Cortés, M. (1998c). A tool for quality service evaluation of the multidisciplinary clinic of lips-clef palate. Proceedings of I Congreso Latinoamericano de Ingeniería Biomédica. Mazatlán, México. pp. 796-799.
53. Ortíz-Posadas, M.R., Vega-Alvarado, L., Jiménez-Jacinto, V., Lazo-Cortés, M., Maya-Behart, J. (1999). Prognosis of cleft palate patients' rehabilitation using a partial precedence algorithm. Proceedings of IV Simposio Iberoamericano de Reconocimiento de Patrones. Conferencia Internacional CIMAF'99. La Habana, pp. 411-418.
54. Ortíz-Posadas, M.R., Vega-Alvarado, L., Maya-Behart, J. (2001). A New Approach to Classify Cleft Lip and Palate. The Cleft Palate-Craniofacial Journal, vol. 38, no. 6, pp. 545–550.
55. Pekalska, E., Duin, R. P. W. (2005). The Dissimilarity Representation for Pattern Recognition. Foundations and Applications. World Scientific.
56. Polczynski, M., & Polczynski, M. (2014). Using the k-Means Clustering Algorithm to Classify Features for Choropleth Maps. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *49*(1), 69-75.
57. Ralambondrainy, H. (1995). A conceptual version of the K-means algorithm. Pattern Recognition Letters, 16, pp. 1147-1157.
58. Rodríguez, A., Egenhofer, M. (2003). Determining Semantic Similarity Among Entity Classes from Different Ontologies. IEEE Transactions on Knowledge and Data Engineering 15(2), pp. 442-456.
59. Rodríguez, A., Egenhofer, M. (2004). Comparing Geospatial Entity Classes: An Asymmetric and Context-Dependent Similarity Measure. International Journal of Geographic Information Science 18(3), pp. 229-256.
60. Ruiz-Shulcloper, J., Fuentes-Rodríguez, A. (1981). Un Modelo Cibernético para el Análisis de la Delincuencia Juvenil. Revista Ciencias Matemáticas Vol. II(1) pp.141-153. Cuba.

61. Ruiz-Shulcloper, J., Pico-Peña, R., Alaminos-Ibarra, C., Valdés-Hernández, G., Manchado-Martín, A. (1992). Modelación Matemática del Problema de Discriminación de Anomalías AGE Perspectivas para Rocas Fosfóricas de Génesis Sedimentaria. Revista Ciencias Matemáticas, vol 13; No.2; pp 159-171. Cuba.
62. Ruiz-Shulcloper, J., Carrasco-Ochoa, J. A., Martínez-Trinidad, J. F. (2013). Reconocimiento de Patrones: Conceptos y Metodología. Reporte Técnico Serie Azul RT_54, CENATAV.
63. Sato, M., Sato, Y. (1995). Extended fuzzy clustering models for asymmetric similarity. In: B. Bouchon-Meunier, R. Yager, and L. Zadeh, (Eds.) Fuzzy Logic and Soft Computing. (Series Advances in Fuzzy Systems-Applications and Theory, 4, World Scientific), pp. 228-237.
64. Sato, Y. (1992). Multidimensional scaling in Minkowski space. Hokkaido Behavioral Science Report, Series M, 20, pp. 69-99.
65. Scheirer W.J., Wilber M.J., Eckmann M., Boult T.E. (2014) Good recognition is non-metric. Pattern Recognition 47(8), 2014, 2721-2731
66. Schölkopf, B., & Smola, A. J. (2002). Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press.
67. Schwering, A., Raubal, M. (2005). Measuring Semantic Similarity between Geospatial Conceptual Regions. Lecture Notes in Computer Science 3799, Springer. 1st International Conference on GeoSpatial Semantics (GeoS 2005), Mexico City, November 29-30, 2005, pp. 90-106.
68. Sebestyen, G.S. (1962). Decision-making process in pattern recognition. ACM Monograph series.
69. Shepard, A. (1979). Adaptive clustering: representation of similarities as combinations of discrete overlapping properties. Psychological Review, 86, (2), 87-123, 1979
70. Tversky, A. (1977). Features of similarity, Psychological Review 84, pp. 327-352.
71. Vapnik, V. N., & Vapnik, V. (1998). Statistical learning theory (Vol. 2). New York: Wiley.
72. Voronin, Yu.A. (1985). Classification Theory and its Application. Editorial Nauka, Moscow.
73. Wilson, D. R., Martinez, T. R. (1997). Improved Heterogeneous Distance Functions, Journal of Artificial Intelligence Research, 6-1, pp. 1-34.
74. Xing Eric P., Ng Andrew Y., Jordan Michael I., Russell Stuart (2002). Distance Metric Learning, with Application to Clustering with Side-information. In Advances in Neural Information Processing Systems 15, Vol. 15, pp. 505-512
75. Xu, R., Wunsch, D.C. (2009). Clustering, John Wiley & Sons, Inc.
76. Ruiz-Shulcloper, J., Abidi (2002). Logical Combinatorial Pattern Recognition: A Review. In: Editor S.G. Pandalai. Recent Research Developments in Pattern Recognition, Pub. Transword Research Networks, USA, 3 pp 133-176.
77. Ruiz-Shulcloper, J. (2008). Pattern Recognition with Mixed and Incomplete Data. Journal Pattern Recognition and Image Analysis, vol. 18, No. 4, pp. 563-576.
78. Martínez-Trinidad, J. Fco., Guzmán-Arenas, A. (2001). The logical combinatorial approach to pattern recognition an overview through selected works. Pattern Recognition 34/4, pp. 741-751.

RT_068, enero 2015
Aprobado por el Consejo Científico CENATAV

Editor: Lic. Lucía González Bayona
Diseño de Portada: Di. Alejandro Pérez Abraham
RNPS No. 2142
ISSN 2072-6287
Indicaciones para los Autores:
Seguir la plantilla que aparece en www.cenatav.co.cu
C E N A T A V
7ma. A No. 21406 e/214 y 216, Rpto. Siboney, Playa;
La Habana. Cuba. C.P. 12200
Impreso en Cuba