

POSIBILIDADES DEL “ANÁLISIS DE ASOCIACIÓN” EN EL ESTUDIO DE LOS EVENTOS EXTREMOS

Gisela del Valle Rodríguez¹, María J. Codornú Pujals¹, Águeda L. García Martín,

¹Departamento de Meteorología. Instituto Superior de Tecnologías y Ciencias Aplicadas. Universidad de La Habana. Cuba. Teléfono-78736157. gisela@instec.cu, marifi@instec.cu, gandon@quimica.cujae.edu.cu

Resumen: En el estudio de los eventos extremos, es cada vez más evidente la conveniencia de incluir la mayor cantidad de información posible sin escatimar las restricciones que el procesamiento habitual exige. Lo anterior conduce a dos problemas: la posible redundancia de la información y la diversidad de la naturaleza de las variables involucradas. Como metodología, el “análisis de asociación” ha permitido identificar con éxito relaciones entre variables en forma de reglas IF-THEN - a partir de datos de diversa tipología (numéricos, nominales, ordinales y otros) cuyas interpretaciones se realizan con ayuda de ciertas métricas asociadas. Los diversos algoritmos desarrollados para estos fines responden a diferentes sistemas conceptuales de índole matemática, destacándose aquellos que posibilitan la obtención de reglas para cualquier tipo de datos como LEM2 y VC-DOMLEM. La aplicación de este último permite distinguir variables de tipo ordinal que frecuentemente están asociadas a los criterios de especialistas y que son particularmente útiles en el ámbito de las investigaciones meteorológicas y climatológicas. En este trabajo se ilustran los resultados de aplicar ambos algoritmos a un caso de inundaciones costeras.

Palabras clave: eventos extremos, minería de datos, análisis de asociación, reglas de decisión.

Abstract: In the study of extreme events, it is increasingly evident the convenience to include as much information as possible, sparing no restrictions that the regular process requires. The above leads to two problems: (i) the potential redundancy of the information, and (ii) the diversity of the nature of the variables involved. Like the methodology, the “association analysis” has allowed successfully to identify the relationship between variables through the rules IF-THEN from data with diverse typology (numeric, nominal, ordinals, etc.), which interpretations are with the assistance of specific metrics related. The various algorithms developed for these purposes, respond to different conceptual systems of mathematical nature. This way highlights those that make possible the extraction of rules for any data, as LEM2 and VC-DOMLEM. The implementation of the latter allows us to distinguish variables of the ordinal kind, which are frequently associated with the specialists’ criteria; they are particularly useful in the meteorological and climate researches. This work illustrates the results to apply both algorithms to a case of severity of coastal floods.

Key Words: extreme events, data mining, association analysis, decision rules

Introducción

El interés por el “análisis de asociación” comenzó desde la década del 90 del pasado siglo, a la vez que aumentaban el tamaño y complejidad de las bases de datos y puede ser entendido en términos de grados de separación entre las transacciones (eventos) independientemente del número de campos (atributos). Es una de las vías del KDD (*Knowledge Discovery in Database*) y que, según su fundador, persigue la “extracción de información implícita, previamente desconocida y potencialmente útil a partir de los datos” (Piatetsky-Shapiro, 1991) a veces coleccionados para otros fines y otras veces intencionalmente creados.

En este caso, son las “reglas de asociación” la expresión del conocimiento simple y útil que se expresa a través de sentencias lógicas IF<antecedentes> THEN <consecuente>, y que fueron conceptualizadas en 1993 por investigadores del *IBM Almaden Research Center* que desarrollaron el algoritmo AIS (Agrawal et al., 1993). AIS realiza diferentes recorridos por la base de datos, va progresivamente creando k-itemsets determinando los relevantes (*large k-itemsets*) e incluyendo los candidatos (*candidate k-itemsets*) y a la vez va utilizando comprobando ciertas medidas de relevancia y confiabilidad. La forma más general de la regla es $X \Rightarrow Y [s, c]$ donde X (atributos de condición) e Y (atributos de decisión) son itemsets disjuntos ($X \cap Y = \emptyset$), s representa la relevancia y c la confiabilidad. Se han definido diversas medidas de interés y existen diferentes criterios en cuanto a su idoneidad para caracterizarlas (Tew et al., 2013) (Ju et al., 2015).

Existe una amplia diversidad de algoritmos para la extracción de reglas, cada uno con sus especificidades, tanto en los fundamentos teórico-metodológicos como en lo operativo. En lo conceptual se destacan los modelos matemáticos que fundamentan los métodos de extracción de reglas, las medidas de interés que garantizan la fiabilidad y relevancia de las mismas, así como la observancia de la naturaleza diversa de los valores de los atributos (cuantitativos, cualitativos y/o semánticos con caracteres numéricos, ordinales, nominales y lingüísticos). En lo operativo, al igual que todo algoritmo, se tienen en cuenta la optimización de los recorridos por las bases de datos, la disminución de los tiempos de ejecución y la trascendencia que en sus procedimientos tiene la tipología de los datos.

No fue hasta inicios de este siglo que se reportaron resultados de aplicación de los algoritmos de inducción de reglas en el ámbito meteorológico y climatológico, siendo mayoritarios aquellos que solamente combinan las formas de escaneo de la base de datos con criterios de medidas de interés realizando un pre-procesamiento de los datos numéricos mediante agrupación por intervalos, por ejemplo:

- Se desarrolló el algoritmo *MOWCATL* (*Minimal Occurrences With Constraints and Time Lags*) para analizar las relaciones entre los parámetros oceánicos y los índices de sequía (Tadesse et al., 2005) encontrándose además los vínculos con el evento ENOS.
- Se concibió una metodología para la predicción puntual de la precipitación a partir de la utilización de un algoritmo de la familia Apriori (*Predictive Apriori*) a datos de una estación meteorológica encontrándose que las reglas obtenidas tienen un valor predictivo de precipitación para 24 horas (Sivaramakrishnan and Meganathan, 2012).

Las reglas de asociación no son de por sí de tipo causal, aunque es posible que adquieran esta significación. Por ejemplo, a partir del conocimiento detallado del sistema físico que causa la generación y progresión de la surgencia de tormenta en cierto sitio del Océano Norte, y de los datos meteorológicos y del nivel de agua en el océano, se creó una base de datos a la cual se le aplicó una modelación lingüística utilizando el algoritmo *linguistic decision tree algorithm* (*LID3*). Las reglas generadas, en forma de árbol,

resultaron útiles y sencillas de comprender constituyendo un elemento predictivo de la surgencia de tormenta para 8 horas y de alerta temprana de inundaciones en esa región (Royston et al., 2013).

El creciente interés en los complejos “eventos extremos” (IPCC, 2014) a la vez ha conducido a bases de datos de mayor dimensionalidad al ser incorporados nuevos campos (atributos) como por ejemplo, los índices ETCCDI de temperatura y precipitación (Klein Tank et al., 2009), los denominados *non-authoritative data* (Schnebele et al., 2014), imágenes de satélite y sus evaluaciones, opiniones de pobladores e incluso otros de índole socioeconómico para los análisis de daños (Done et al., 2018) (Beken and Eminagaoglu, 2019). En este contexto, dejan de ser adecuados los algoritmos de inducción de reglas basados en los criterios de medidas pues se generan con mayor frecuencia:

- Bases de datos en las que, con diferentes valores de atributos las transacciones pertenecen a la misma clase, es decir, presentan “indiscernibilidad”
- Bases de datos con transacciones que no pueden describirse de forma precisa, ni pertenecientes a un conjunto ni a su complemento, es decir, se caracterizan por la “ambigüedad o vaguedad”.
- Bases de datos “no consistentes”, con transacciones con iguales valores de atributos de condición que poseen diferentes clases o valores de atributos de decisión.

Resultado de una extensión de la Teoría de Conjuntos, y a partir del trabajo fundacional de Pawlak, se desarrolló una teoría matemática (Pawlak, 1981) (Pawlak and Skowron, 2007) (Yao, 2007) (Yao, 2016) que está basada en:

- La “indiscernibilidad” como relación que permite tratar las transacciones que no pueden ser caracterizadas de manera precisa en términos de valores de sus atributos.
- La “aproximación” de las transacciones en tres categorías: aproximación alta, aproximación baja y la región frontera, esta última cuando pueden pertenecer o no a las dos primeras.
- La “inconsistencia” de la base de datos puede ser tratada teniendo en cuenta, además de las dos propiedades anteriores, la “similaridad”.

Estas propiedades constituyen los pilares de una teoría matemática, con sus correspondientes modelos y métodos, que se ha consolidado (Skowron and Dutta, 2018) fundamentando otra manera de enfocar el tratamiento de datos. Los algoritmos también han evolucionado incorporando al primario LEM2 (Slowinski and Vanderpositen, 1997) algunas modificaciones y extensiones sin variar la esencia del modelo básico (Pawlak, 2000). De manera especial, la profundización en la propiedad “dominancia”, cuya observancia por encima de la “indiscernibilidad” permite distinguir de forma específica la ordinalidad de algunos atributos, requiere de otros algoritmos como DOMLEM y VC-DOMLEM (Slowinski et al., 2012).

La apropiación de esta otra manera de inducir las reglas ha sido aplicada en el ámbito meteorológico y climatológico en estudios sobre “eventos extremos” como la clasificación de datos volumétricos de radar meteorológico utilizados para detectar eventos de tormenta responsables de tiempo severo (Peters et al., 2003), la determinación de tipos de nubes preferentemente asociadas a la génesis de tormentas severas (Chaudhuri, 2008), la obtención de reglas espacio-temporales en el estudio de típicos torbellinos mesoescalares (Du et al., 2011), el estudio de la contaminación del aire y su relación con extremos de precipitación (Payus et al., 2012), entre otros.

A través de la aplicación de los algoritmos LEM2 y VC-DOMLEM a una misma base de datos, las autoras ilustran las posibilidades del “análisis de asociación” y la importancia de distinguir los campos (atributos) con valores regulares de aquellos que deban ser caracterizados por dominios ordenados.

Materiales y métodos

Aspectos tecnológicos

Un sistema de información está formado por los objetos a describir y sus atributos. Los objetos a describir se colocan en filas y los atributos de condición y el (los) atributo(s) de decisión se ordenan en columnas formándose así la tabla de decisión.

Los algoritmos LEM2 y VC-DOMLEM están diseñados para encontrar vínculos entre los campos(atributos) y las transacciones (objetos)de un sistema de información. Una peculiaridad común a ambos es que no se requiere realizar ningún pre-procesamiento inicial a los datos.

Fueron seleccionadas dos aplicaciones informáticas ROSE2 y jMAF, desarrolladas ambas en la Universidad de Poznan, Polonia.

- **ROSE2** (*Rough Set Data Explorer*). Es un software en el cual se implementan los elementos básicos de la *RST (Rough Set Theory)* y las técnicas de descubrir reglas. Ha sido creado en el Laboratorio de Sistemas de Soporte a las Decisiones Inteligentes del Instituto de Ciencia Computacional en Polonia, basado en cuarenta años de experiencia en *Rough Sets* para el análisis de las decisiones y descubrimiento del conocimiento. El sistema ROSE2 es sucesor de los sistemas *RoughDAS* y *RoughClass*. *RoughDAS* y es desde el punto de vista histórico la primera implementación exitosa de la *RST*, la cual ha sido utilizada en muchas aplicaciones de la vida real (Slowinski et al., 1999). Tiene implementado tres modelos RST de exploración de datos, obtención de reductos, ofrece cuatro variantes y tres esquemas de inducción de reglas, todo ello con el riguroso respaldo(Predki et al., 1998).

Referente actual: <http://www-idss.cs.put.poznan.pl/rose> .

Prof. Roman Slowiński. Instituto de Matemáticas. Universidad de Poznan. Poznan. Polonia.

- **jMAF**. Es un marco de trabajo para el análisis de datos destinado a resolver problemas de clasificación con criterios múltiples basado en los *roughs sets* (conjuntos aproximados) con relación de dominancia, también denominados problemas de clasificación ordinal con restricciones monotónicas(Blaszczyński et al., 2009). Las dos metodologías de razonamiento sobre los datos (Slowinski et al., 2012): *DRSA-Dominance-based Rough Set Approach* y *VC-Variable Consistency Dominance-based Rough Set Approach* dan nombre a los dos métodos de clasificación (DRSA y VCDRSA). Para la extracción de reglas está implementado el algoritmo VC-DOMLEM, que garantiza un adecuado tratamiento a problemas con atributos regulares o con dominios ordenados. Desarrollado en Lenguaje JAVA, está basado en las librerías java Rough Set (jRS). Sucesor de 4eMka y JAMM, tiene versiones para Windows y para Linux. Ha sido desarrollado por los grupos jRS y jMAF de la Universidad de Tecnología de Poznan.

Referente actual: <http://www.cs.put.poznan.pl/jblaszczyński/Site/jRS.html>

Prof. Jerzy Blaszczysky. Instituto de Ciencias de la Computación. Universidad de Tecnología de Poznan. Poznan. Polonia.

Los dos algoritmos aplicados para extraer las reglas de asociación fueron LEM2 y VC-DOMLEM. Con LEM2 todas las variables pueden ser declaradas regulares y solamente con VC-DOMLEM es necesario

distinguir aquellas respecto a las cuales se requiere de declarar los dominios en orden creciente o decreciente. Otras especificidades se muestran en la Tabla 1.

Tabla 1. Características distintivas de las aplicaciones informáticas utilizadas

| ASPECTOS | ROSE2 | jMAF |
|-------------------------|--|--|
| Teoría | <i>RST-Rough Set Theory</i> | <i>RST-Rough Set Theory</i> |
| Lenguaje | C++ | JAVA |
| Sistema Operativo | MS Windows | MS Windows/Mac OS, LINUX |
| Interface de usuario | GUI | GUI |
| Licencia | No lucrativo | No lucrativo |
| Conceptos básicos | SI | SI |
| Discretización | SI | NO |
| Selección de atributos | SI | SI |
| Selección de instancias | NO | NO |
| Valores perdidos | SI | NO |
| Validación cruzada | <i>k-fold cross-validation/ leaving-one-out</i> | VC-DOMLEM |
| Inducción de reglas | LEM2 | VC-DOMLEM |
| Relaciones | Equivalencia/Similaridad | Dominancia |
| Propiedades | <i>RST-Classical</i> (Pawlak,1982)/ <i>VPRST-Variable Precision Rough Set Model</i> (Ziarko, 1993) | <i>RST-VCDRSA</i> |
| Tipología de reglas | exactas/aproximadas | seguras/posibles/aproximadas certeza/posibilidad/duda |
| Significado | para una clase /para más de una clase | al menos/ a lo más |
| Clasificación | SI | SI |
| Reclasificación | NO | <i>DRSA method</i> <i>VCDRSA method</i> |

Interpretación de las reglas

La interpretación de una regla de decisión depende de la tipología de la regla y de las medidas de interés.

Tipología de las reglas de asociación

ROSE2. Contempla dos algoritmos, el *RST-Classical* de Pawlak (Pawlak, 1982) y el *VPRST-Variable Precision Rough Set Model* desarrollado por Ziarko y colaboradores.(Ziarko, 1993). En

ambos casos las estructuras de las reglas son IF < antecedentes> THEN <consecuente> y

pueden ser exactas (con solo una clase de decisión) o aproximadas (cuando pueden ser más

de una clase de decisión).

jMAF. Con el algoritmo *VC-DOMLEM* se extraen reglas que tienen en cuenta el primer lugar el grado de certeza o seguridad (*CERTAIN*) y en segundo lugar el grado de posibilidad (*POSSIBLE*) pero que además poseen otro nivel de gradación de la clase de decisión, es decir, “a lo más” (“como máximo” o “cuando más”) y “al menos” (“como mínimo” o “cuando menos”). Se tendrán cinco posibles significados: *CERTAIN AT_LEAST*, *CERTAIN AT_MOST*, *POSSIBLE AT_LEAST*, *POSSIBLE AT_MOST* y una última variante en situación de duda.

Medidas de interés

Las medidas de interés pueden calcularse a partir de la relación entre cardinalidades de ciertas combinaciones en la regla $C \rightarrow_x D$.

- En el sistema de información y la tabla de decisión, el concepto de cardinalidad de los conjuntos involucrados $S = (U, C, D)$ se denota por $card(U)$ y significa el total de objetos del universo, así como $card(C)$ y $card(D)$ corresponden al número de objetos que satisfacen ciertos valores de los atributos de condición y decisión, respectivamente.
- Para el caso de la regla $C \rightarrow_x D$ se denota como $card(C \cap D)$ al número de objetos que satisfacen la parte de la condición de la regla (cubiertos por la regla) y que pertenecen a la clase de decisión sugerida.

Para caracterizar las reglas, las medidas de interés más utilizadas son:

- Soporte de una regla. Es la cardinalidad de la regla denotada por $supp(C, D) = card(C \cap D)$ que expresa el número de objetos que en la tabla de decisión respaldan la regla $C \rightarrow_x D$ y se denomina soporte sup- [*support*].
- Fortaleza, certeza y cobertura de una regla. Son ratios que se obtienen de dividir el soporte de la regla entre las cardinalidades de los conjuntos U, C y D respectivamente por lo que se obtiene, en ese orden, la fortaleza str- [*strength*], la certeza o confianza conconf- [*confidence*] y la cobertura cov- [*coverage*]. La fortaleza se interpreta como la representatividad de la regla en la tabla de decisión, [*confidence*] como el grado de confianza o certeza de la regla y [*coverage*] como el grado de confianza de la regla inversa o las razones de la decisión.
- La medida que cuantifica la “inconsistencia” incon- [*InconsistenceMeasure*] y la información que brinda la existencia de ejemplos negativos justifica la disminución del cov- [*coverage*].

En dependencia del objetivo del estudio, pueden llegar a ser relevantes otras que destacan otras aristas del “conocimiento aproximado” y que pueden ser muy útiles para su interpretación. Además, la calidad de las aproximaciones puede ser cuantificada en varias formas. La precisión cuantifica la completitud del conocimiento y las medidas de la granularidad cuantifican la resolución del conocimiento (Yao, 2010).

Aspectos metodológicos

Caso estudio: Base de datos de inundaciones costeras

La base de datos utilizada fue creada para un estudio, desarrollado en el año 2013, sobre inundaciones costeras causadas por sistemas frontales ocurridas en el litoral habanero en el período 1980-2010 (Hernández, 2013).

En la Tabla 2 se muestran los atributos y los valores de cada uno además de señalarse la categoría numérica del dato asociado. La Tabla 2 muestra el sistema de variables especificándose los valores o categorías de cada una según su naturaleza cuantitativa o cualitativa, numérica por intervalos o nominales, distinguiendo aquellas que son ordinales con dominios ordenados (VmVBoya, HsigBoya e IndENOS).

La intensidad de la inundación fue revisada atendiendo a los criterios referidos en el Manual de Procedimientos de Meteorología Marina del INSMET (INSMET, 2015) relativos a la altura significativa de la ola y la conciliación con la cronología de inundaciones costeras confeccionada por Hidalgo y colaboradores (Hidalgo et al., 2016). En total fueron 39 inundaciones previamente clasificadas en 26 ligeras, 9 moderadas y 4 fuertes.

Se dispuso de dos variables relativas a la boya 42003 NDBC, NOAA (26,044 LN y 85,612 LW) del Golfo de México. Tanto para la altura significativa de la ola (HsigBoya) como para la velocidad media del viento (VmVBoya), los dominios de estas dos variables coincidieron con los criterios que aparecen en el Manual de Procedimientos de Meteorología Marina del INSMET (INSMET, 2015).

Respecto al Índice ENOS (IndENOS) como indicador adimensional (calculado a partir de la media trimestral de los últimos tres meses de la anomalía de la temperatura superficial del mar y la media de estos tres meses del índice de Oscilación del Sur) fueron establecidas cinco categorías: N-neutro, D-débil, M-moderado, F-fuerte y MF-muy fuerte (Hernández, 2013).

Tabla 2. Caracterización de las variables seleccionadas para el estudio de las inundaciones costeras en el litoral habanero (2000-2010).

| ATRIBUTOS | | VALORES - ATRIBUTOS | DATO |
|---|----------|--|---------------------|
| Situación meteorológica | SitMET | {Frente Frío, Baja Extratropical} {FF,B} | nominal |
| Tipología del sistema frontal | Tipo | {Clásico, Reversino, Secundario } {C, R, S} | nominal |
| Velocidad media del viento en la boya 42003 [km/h] | VmVBoya | { ≤ 9.7 , (9.7, 15.3)], ≥ 15.3 } { menor9.7, media, mayor15.3 } | continua ordinal |
| Dirección predominante del viento | DirPred | { W, NW, NNW, NE, NNE } | nominal |
| Altura significativa de la ola en la boya 42003 [m] | HsigBoya | { < 4, [4, 5], > 5 } { menor4, media, mayor5 } | continua ordinal |
| Índice ENOS | IndENOS | {<48, [48,150), [150,350), [350.800], >800} {N, D, M, F, MF} | continua ordinal |
| Intensidad de la inundación | INUND | {LIGERA, MODERADA, FUERTE} | nominal |

Resultados y discusión

Fueron extraídas las reglas de asociación para cada algoritmo aplicado a la misma base de datos. Las reglas de asociación obtenidas se han representado en forma de tabla. Primeramente, se muestra en la Tabla 3 las reglas de asociación obtenidas con VC-DOMLEM y con la misma estructura de tabla, intencionalmente se han colocado las reglas obtenidas con LEM2 en la Tabla 4.

El conocimiento del fenómeno bajo estudio, desde la perspectiva del “análisis de asociación”, se resume en lo mostrado en las tablas. Intencionalmente se ha utilizado una misma estructura y como era de esperar, existe mayor detalle de información en las reglas cuya extracción se basa en la “dominancia” (VC-DOMLEM) respecto a la información en las reglas derivadas del algoritmo que se basa en la “similaridad” (LEM2). Las diferencias no solamente se manifiestan en las medidas asociadas a las reglas que son necesarias para su interpretación, sino en la tipología de las mismas.

Como punto de partida, debe insistirse en que son dos enfoques diferentes, ninguno superior al otro, aunque la posibilidad de ordenar aquellos atributos de tipo ordinal concede notoriedad a los resultados obtenidos con VC-DOMLEM.

Serán destacados aquellos aspectos de interés general, lo cual corresponde con los objetivos de este trabajo.

Generalidad de las reglas de asociación

- En ambas tablas existen reglas de diferentes longitudes, es decir, cantidad de atributos de condición. Aquellas que son más largas, son menos generales, por lo que las que tienen tres atributos de condición involucrados serán menos generales que las que poseen dos, las cuales están en mayoría.
- No obstante, las reglas de un solo atributo de condición merecen un análisis aparte, por cuanto no representan una información relevante. Estas reglas de un solo atributo de condición tienen una connotación descriptiva y puede verificarse que corresponden plenamente con los resultados de aplicar la Estadística Descriptiva como es el caso del estudio original de la base de datos desarrollada en el 2013 (Hernández, 2013).
- Aquellas reglas cuya medida de inconsistencia sea no nula están concentradas en el enfoque basado en la “dominancia” lo cual está respaldado por los ejemplos negativos que hacen disminuir el *cov [coverage]* y debilitan las razones de la decisión (clase o atributo de decisión).

Significatividad de los atributos en las reglas

La frecuencia de los atributos de condición en las reglas constituye una medida de su significatividad y se encuentra reflejada en los histogramas de la Figura 1.

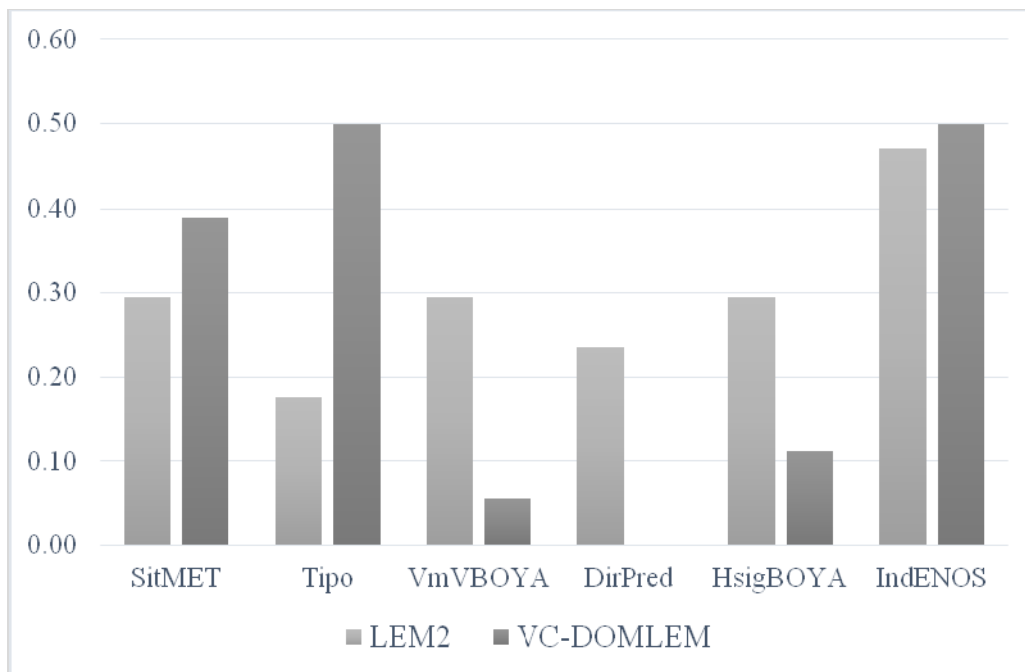


Figura 1. Significatividad comparativa de los atributos de condición en las reglas de asociación extraídas a través de los algoritmos LEM2 y VC-DOMLEM

Es necesario aclarar que la significatividad en las reglas no indica independencia del fenómeno respecto a los atributos, sino su importancia de la presencia de éste en la comprensión de los vínculos con otros atributos. Con esta perspectiva pueden hacerse las siguientes observaciones:

- Con LEM2 la significatividad los atributos no es tan diferente siendo la del atributo IndENOS la más importante.
- Con VC-DOMLEM se incrementa la significatividad de los atributos característicos SitMET(frente frío, baja extratropical) y Tipo (clásico, reversino, secundario) siendo destacado el reforzamiento de IndENOS.
- Es curioso que, en contraposición a lo anterior, las otras dos variables de tipo ordinal, VmVBoya, y HsigBoya, al igual que la otra característica del viento DirPred disminuyan sensiblemente su presencia en las reglas de asociación.

Proceso de reclasificación

La aplicación jMAF permite conciliar la validez de las reglas con la clasificación inicial que trae la propia base de datos. El proceso de reclasificación resultó de un 79,48% de efectividad y 20,51% incorrecto. La matriz confusión destaca que las inundaciones fuertes y las ligeras tienen una máxima calidad de clasificación, pero en el caso de las moderadas solamente un caso fue confirmado existiendo 7 clasificadas de ligera y solo una de fuerte.

En la Tabla 5 se ha seleccionado un caso en el cual al reclasificar se eleva su categoría junto a las reglas que lo fundamentan y en la Tabla 6 se muestran los resultados de un caso en que la reclasificación arrojó una clasificación disminuida en su categoría de decisión. Para este análisis se requiere tener en cuenta además de la tipología de la regla, la tipología de los valores de los atributos que toman valores según la relación “<=” o según la opuesta “>=”.

Tabla 5. Reglas que explican el caso de un objeto (evento) del cual se modifica su clase de decisión a una categoría superior

| | Ejemplo: Objeto 24 Decisión original: MODERADA Resultado de clasificación: FUERTE Reglas <i>matcheadas</i> : 5 | Tipología de la regla |
|---|---|-----------------------|
| 1 | (SitMET=B) & (HsigBoya) >= [4,5] ⇒ (INUND<=MODERADA) | CERTAIN AT_LEAST |
| 2 | (SitMET=B) & (IndENOS<= M) ⇒ (INUND<=MODERADA) | CERTAIN AT_LEAST |
| 3 | (Tipología=S) ⇒ (INUND>=MODERADA) | CERTAIN AT_MOST |
| 4 | (IndENOS<=F) ⇒ (INUND<=MODERADA) | CERTAIN AT_LEAST |
| 5 | (Tipología=S) ⇒ (INUND>=MODERADA) | CERTAIN AT_MOST |

Tabla 6. Reglas que explican el caso de un objeto (evento) del cual se modifica su clase de decisión a una categoría inferior

| | Ejemplo: Objeto 23 Decisión original: MODERADA Resultado de clasificación: LIGERA Reglas <i>matcheadas</i> : 4 | Tipología de la regla |
|---|---|-----------------------|
| 1 | (SitMET=FF) ⇒ (INUND>=MODERADA) | CERTAIN AT_LEAST |
| 2 | (IndENOS<= F) ⇒ (INUND<=MODERADA) | POSSIBLE AT_LEAST |
| 3 | (Tipología=C) & (VmVBoya>=[9,7. 15.3]) ⇒ (INUND<=MODERADA) | CERTAIN AT_MOST |
| 4 | (IndENOS<=F) ⇒ (INUND>=LIGERA) | POSSIBLE AT_LEAST |

Existen otros análisis que pueden realizarse profundizando en la interpretación de los vínculos desde el punto de vista de la especialidad, pero en este trabajo solamente se aspira a mostrar las posibilidades del “análisis de asociación”.

Conclusiones

El “análisis de asociación” puede contribuir a la mejor comprensión de los fenómenos en los que se manifiesta inconsistencias en los valores de sus atributos.

La selección del algoritmo-DOMLEM permite distinguir la relación de “dominancia” tanto en las clasificaciones o valores de la clase de decisión que dentro de las interrelaciones entre los valores de sus diferentes atributos.

En fenómenos donde la discretización de valores continuos de variables meteorológicas requieren de criterios de especialistas, la perspectiva de la incertidumbre desde la “relación de dominancia” podría resultar provechosa.

Recomendaciones y agradecimientos

Las autoras de este trabajo desean agradecer a la Dra. Ida Mitrani Arenal su colaboración en los aspectos técnicos del trabajo.

Referencias bibliográficas

- Agrawal, R., Imielinski, T., Swami, A., 1993. Mining Association Rules between Sets of Items in Large Databases, in: Proceedings of the 1993 ACM SIGMOD Conference. Presented at the International Conference of Management of the Data, Washington DC, USA, pp. 207–216.
- Beken, S., Eminagaoglu, M., 2019. An Information Security Risk Assessment Model Based on Bayesian Network and Fuzzy Inference System. *Ege Strat. Arařtırmalar Derg.* 10, 13–33.
- Blaszczynski, J., Greco, S., Matarazzo, B., Slowinski, R., Szelag, M., 2009. jMAF-Dominance-based Rough Set Approach Data Analysis Framework. User’s guide, Rough Set Theory-Software. Poznan University, Poznan, Poland.
- Chaudhuri, S., 2008. Preferred type of cloud in the genesis of severe thunderstorms. A soft computing approach. *Atmospheric Res.* 88, 149–156.
- Done, J.M., Pai Mazumder, D., Towler, E., Kishtawal, C.M., 2018. Estimating impacts of North Atlantic tropical cyclones using an index of damage potential. *Clim. Change* 146, 561–573.
- Du, Y., Fan, X., He, Z., Su, F., Zhou, C., Mao, H., Wang, D., 2011. Extraction of spatial-temporal rules from mesoscale eddies in the South China Sea based on rough set theory. *Ocean Sci.* 7, 835–849. doi:10.5194/os-7-835-2011
- Hernández, I., 2013. Las inundaciones costeras generadas por sistemas frontales en el malec3n habanero (Trabajo de Diploma-Licenciatura en Meteorología). Instituto Superior de Tecnologías y Ciencias Aplicadas, La Habana, Cuba.
- Hidalgo, A., Pérez, G., Mitrani, I., Hernández, N., C3rdova, O.L., Regueira, V., Ram3rez, W., Rodr3guez, C.M., 2016. Cronolog3a de las inundaciones costeras en Cuba-Programa Meteorolog3a y Desarrollo Sostenible del Pa3s (Proyecto “Procedimiento para la ejecuci3n y uso de las observaciones del estado de la superficie marina desde estaciones costeras, en la predicci3n de oleaje e inundaciones costeras en territorio cubano” No. P211LH007-015). Instituto de Meteorolog3a, La Habana, Cuba.
- INSMET, 2015. Manual de Procedimientos de Meteorolog3a Marina. INSMET, La Habana, Cuba.
- IPCC, 2014. Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press-United Kingdom and New York, NY, USA.

- Ju, C., Bao, F., Xu, C., Fu, X., 2015. A Novel Method of Interestingness Measures for Association Rules Mining Based on Profit. *Discrete Dyn. Nat. Soc.* 2015, 10.
- Klein Tank, A.M.G., Zwiers, F.W., Zhang, X., 2009. Guidelines on Analysis of extremes in a changing climate in support of informed decisions for adaptation, Climate Data and Monitoring-WCDMP-No.72-WMO-TD No. 1500. World Meteorological Organization.
- Pawlak, Z., 2000. *Rough Sets*, Polish Academy of Sciences. ed. University of Information Technology and Management, Warsaw, Poland.
- Pawlak, Z., 1982. *Rough Sets*. *Int. J. Comput. Inf. Sci.* 11, 341–356.
- Pawlak, Z., 1981. *Classification of Objects by Means of Attributes* (Reports No. 429). Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland.
- Pawlak, Z., Skowron, A., 2007. *Rough Sets: Some extensions*. *Inf. Sci.* 177, 28–40.
- Payus, C., Sulaiman, N., Bakar, A.A., 2012. Influence of Air Quality towards the Extreme Rainfall Events: Rough Set Theory. *Int. J. Eng. Res. Appl. IJERA* 2, 544–550.
- Peters, J.F., Suraj, Z., Shan, S., Ramanna, S., Pedrycz, W., Pizzi, N., 2003. Classification of meteorological volumetric radar data using rough set methods. *Pattern Recognit. Lett.* 24, 911–920.
- Piatetsky-Shapiro, G., 1991. Knowledge Discovery in Real Databases. *Artif. Intell. Mag.* Winter.
- Predki, B., Slowinski, R., Stefanowski, J., Susmaga, R., Wilk, S., 1998. Rough Set Data Explorer-ROSE-Software implementation of the rough set theory, in: Polkowski, L., Skowron, A. (Eds.), *Lecture Notes in Artificial Intelligence*. Presented at the Rough Sets and Current Trends in Computing'98 Conference, Springer Berlin, pp. 605–608.
- Royston, S., Lawry, J., Horsburgh, K., 2013. A linguistic decision tree approach to predicting storm surge. *Fuzzy Sets Syst.* 215, 90–111. doi:10.1016/j.fss.2012.10.001
- Schnebele, E., Cervone, G., Waters, N., 2014. Road assessment after flood events using non-authoritative data. *Nat. Hazards Earth Syst. Sci.* 14, 1007–1015. doi:10.5194/nhess-14-1007-2014
- Sivaramakrishnan, T.R., Meganathan, S., 2012. Point Rainfall Prediction using Data Mining Technique. *Res. J. Appl. Sci. Eng. Technol.* 4, 1899–1902.
- Skowron, A., Dutta, S., 2018. Rough sets: past, present, and future. *Nat. Comput.* 17, 855–876. doi:10.1007/s11047-018-9700-3
- Slowinski, K., Slowinski, R., Stefanowski, J., Fibak, J., 1999. *ROSE2. Rough Set Data Explorer-User's Guide*.
- Slowinski, R., Greco, S., Matarazzo, B., 2012. Rough Set and Rule-Based Multicriteria Decision Aiding. *Pesqui. Oper.* 32, 213–269.
- Slowinski, R., Vanderpooten, D., 1997. Similarity relation as a basis for rough approximations, in: Wang, P.P. (Ed.), *Advances in Machine Intelligence and Soft Computing*. Bookwrights Raleigh, pp. 17–33.
- Tadesse, T., Wilhite, D.A., Hayes, M.J., Goddard, S., 2005. Discovering Associations between Climatic and Oceanic Parameters to Monitor Drought in Nebraska Using Data-Mining Techniques.
- Tew, C., Giraud-Carrier, C., Tanner, K., Burton, S., 2013. Behavior-based clustering and analysis of interestingness measures for association rule mining. *Data Min. Knowl. Discov.* 27. doi:10.1007/s10618-013-0326-x
- Yao, Y., 2016. Rough-set concept analysis: Interpreting RS-definable concepts based on ideas from formal concept analysis. *Inf. Sci.* 346-347, 442–462.
- Yao, Y.Y., 2010. Notes on Rough Set Approximations and Associated Measures. *J. Zhejiang Ocean Univ.* 29, 399–410.
- Yao, Y.Y., 2007. *Decision-Theoretic Rough Set Models*. Presented at the Second International Conference, *Rough Sets and Knowledge Technology-2007*, pp. 1–12.
- Ziarko, W., 1993. Variable Precision Rough Set Model. *J. Comput. Syst. Sci.* 46, 39–59.