Método de detección no determinista de *outliers* basado en el modelo de conjuntos aproximados de precisión variable

Alberto Fernández-Oliva[1], Miguel Abreu-Ortega[1], Covadonga Fernández-Baizán[2], Francisco Maciá-Pérez[3]

[1]Departamento de Ciencia de la Computación, Facultad de Matemática y Computación
Universidad de la Habana
afdez@matcom.uh.cu, miguel87@lab.matcom.uh.cu

[2]Departamento de Lenguajes y Sistemas Informáticos e Ingeniería del Software
Facultad de Informática de la Universidad Politécnica de Madrid
cfbaizan@fi.upm.es

[3]Departamento de Tecnología Informática y Computación
Universidad de Alicante
pmacia@dtic.ua.es

Resumen. En este trabajo se presenta un procedimiento para la detección de outliers basado en el modelo de conjuntos aproximados de precisión variable (*Variable Precision Rough Sets Model —VPRSM*) que tiene como esencia una generalización del concepto estándar de la relación de inclusión de conjuntos definido en el modelo básico de conjuntos aproximados *(Rough Sets Basic Model —RSBM)*. Dicha extensión propone una clasificación con un cierto grado de incertidumbre con el objetivo de mejorar la calidad de la detección. A partir del método se propone un algoritmo computacionalmente eficiente y se presentan los resultados alcanzados tras aplicar el método a un caso real y comparando los mismos con los alcanzados mediante el algoritmo basado en *RSBM*.

1 Introducción

La detección de casos excepcionales (*outlier detection*) es un campo de creciente relevancia dentro del más general de *Data Mining*. Si en éste último el objetivo es extraer patrones de conocimiento a partir de grandes volúmenes de datos lo más generales posible —expresión de tendencias que ignoran por tanto la marginalidad o excepción—, en *outlier detection* se adopta el punto de vista opuesto, lo que puede reportar hallazgos de conocimiento de suma importancia estratégica en una variada gama de aplicaciones: detección de fraudes, detección de accesos ilegales a redes corporativas, detección de errores en datos de entrada, etc.

El modelo básico de conjuntos aproximados *(Rough Sets Basic Model —RSBM)* propuesto por el profesor Z. Pawlak [1] en 1982 es un modelo con una base matemática simple y sólida que parte de la teoría de conjuntos aproximados *(Rough Sets —RS)*: la teoría de relaciones de equivalencia que aquí permite describir

particiones constituidas por clases de objetos indiscernibles. En los últimos años este modelo se ha aplicado exitosamente en diversos contextos tales como *Knowledge Discovery from Data*, *Data Mining*, *Machine Learning*, *Expert Systems, Decision Support Systems*, etc. Por tal motivo, el estudio del mismo ha acaparado la atención de los académicos e investigadores a nivel internacional.

En Jiang et. al [2] se propone un planteamiento alternativo sobre dicho modelo que constituye un nuevo enfoque del problema de *outliers detection* basado en *RSBM.* Bajo este enfoque, los *outliers* se definen como elementos de los *conjuntos excepcionales no redundantes* que poseen un *grado de marginalidad* mayor que un umbral establecido. Si bien la idea subyacente resulta intuitiva, deviene en un problema intratable por ser de orden exponencial. Este problema se soluciona gracias a una extensión del marco teórico para construir un algoritmo de detección de *outliers* basado en *RSBM* con orden de complejidad temporal y espacial no exponencial.

A pesar de esta solución, debido a la concepción de *RSBM,* sigue sin ser posible modelar problemas y situaciones donde sea necesario permitir la clasificación con un grado controlado de incertidumbre o un posible error de clasificación. En la práctica, poder admitir algún nivel de incertidumbre en el proceso de clasificación puede llevar a una comprensión más profunda y a una mejor utilización de las propiedades de los datos analizados. En tal sentido, un algoritmo propuesto a partir de *RSBM* hereda esta limitación debido a una rigurosa forma de modelar la inclusión de conjuntos.

Una generalización de *RSBM* es la propuesta por W.Ziarko, el modelo de conjuntos aproximados de precisión variable (*Variable Precision Rough Sets Model —VPRSM*) [3]. Este modelo subsana el carácter determinista (en lo que respecta a la clasificación) que presenta el planteamiento de Jiang et. al [2] partiendo de una idea muy simple, la *relajación* del concepto de inclusión de conjuntos, permitiendo así manejar unos umbrales definidos por el usuario.

En objetivo principal de este trabajo es crear un método de detección no determinista de *outliers* basado en *VPRSM* computacionalmente viable partiendo de la hipótesis de que el modelo *VPRSM* permite ampliar la aplicación del método original basado en *RSBM* a contextos en los que sea necesaria una clasificación con un cierto grado de incertidumbre. Teniendo en cuenta esto, el resto del artículo se ha estructurado de la siguiente forma: en el apartado 2 se realiza una comparativa entre los modelos *RSBM* y *VPRSM*, destacándose algunas de las limitaciones que se le señalan a *RSBM* y que sirven de antecedente para la concepción del modelo *VPRSM*; en el apartado 3 se exponen las notaciones básicas y las propiedades esenciales que caracterizan a *VPRSM* y que constituyen el marco teórico del nuevo procedimiento de detección que se propone; en el apartado 4 se completa el marco teórico propuesto con el diseño y la instrumentación computacional del algoritmo de detección; la funcionalidad del mismo se ilustra con un ejemplo en el apartado 5; en el apartado 6 se establecen comparaciones en cuanto a tiempo de ejecución y calidad en la detección entre el algoritmo basado en *RSBM* y el basado en *VPRSM*; finalmente, en el apartado 7 se presentan las principales conclusiones que se derivan de esta investigación y se plantean las líneas futuras a seguir.

2 RSBM frente a VPRSM

La minería de datos (*Data Mining*) emerge cada vez con más fuerza como un área de la Inteligencia Artificial que brinda técnicas, teorías y herramientas para el análisis de datos en los complejos *data sets* de hoy día. La Teoría de *RS* en este ámbito incide también decisivamente en el empeño por alcanzar tales metas. Desde finales de la década de los 80 ya se reportan resultados importantes de la aplicación de la misma y en los últimos años, como ya hemos señalado, su aplicación en múltiples contextos investigativos, especialmente en los procesos de *Knowledge Discovery on Data (KDD) – Data Mining,* pone de manifiesto su efectividad en la solución de disímiles problemas. Todo ello demuestra su versatilidad y sus diversos entornos de aplicación. Por solo citar algunos ejemplos, pueden señalarse los siguientes: en el ámbito de los Sistemas de Mercado (*Trading Systems*) [4], en diferentes aplicaciones referidas al *Machine Learning* [5], en investigaciones básicas orientadas al desarrollo de Sistemas Inteligentes [6] [7], en el campo de la Bioinformática [8], etc.

El problema fundamental de la teoría de *RS* [9] es facilitar el análisis de la clasificación. La aproximación (superior e inferior) se hace necesaria ante la incapacidad de establecer, con el conocimiento disponible, clasificaciones completas de objetos que pertenecen a una cierta categoría.

Con cierta frecuencia, la información disponible permite sólo hacer clasificaciones parciales y, en tales casos, la teoría de *RS* puede utilizarse con efectividad para modelar este tipo de clasificación pero, a partir de esta teoría, dicha clasificación debe ser completamente correcta o cierta. Esto limita la posibilidad de concebir una clasificación con un grado controlado de incertidumbre, es decir, la posibilidad de que exista un cierto error en la clasificación. Esto está fuera de la realidad en *RSBM*. Paradójicamente, en la práctica y en muchos casos, resulta conveniente admitir algún nivel de incertidumbre en el proceso de clasificación, lo cual puede permitir una mejor comprensión y utilización de las propiedades de los datos que se están analizando.

Otra limitación señalada a *RSBM* es que asume que el universo *U* de objetos o datos tenidos en consideración es conocido y que todas las conclusiones derivadas de la aplicación del referido modelo son aplicables solamente a ese conjunto de objetos. Sin embargo, en la práctica, hay necesidad de generalizar las conclusiones obtenidas a partir de un pequeño conjunto de objetos (*U*) hacia un universo mayor, por ejemplo, el mundo real.

RSBM permite obtener hipótesis basadas sólo en reglas de clasificación libres de errores (las cuales se expresan en la aproximación inferior, $\underline{X}$) que se obtienen del análisis de los datos tenidos en consideración (*U*), es decir, este modelo es determinista. Sin embargo, hay múltiples situaciones en el mundo real que avalan la necesidad de tener también en cuenta clasificaciones parcialmente incorrectas. Una regla de clasificación parcialmente incorrecta proporciona también información útil. Puede establecer la tendencia de los valores si la mayoría de los datos disponibles a los que se aplica la regla pueden clasificarse correctamente. Precisamente, *VPRSM* brinda la posibilidad de detectar o establecer esta tendencia de la información y, a partir de ella, realizar determinados análisis sobre un cierto universo de objetos o datos, es decir, es un modelo estadístico [10].

En el siguiente apartado se destacan los aspectos más relevantes del modelo de precisión variable los cuales se orientan fundamentalmente, a resolver las limitaciones señaladas a *RSBM*. La esencia radica en una nueva concepción o generalización del concepto estándar de relación de inclusión de conjuntos.

3 *VPRSM*. Notaciones básicas y propiedades

VPRSM es una generalización de *RSBM*. Se deriva del mismo, sin asumir adicionalmente nada más. A partir de esta generalización se permite el manejo de información con un cierto grado de incertidumbre. Igualmente, se reportan diversos resultados investigativos relacionadas con la aplicación de este modelo. Por ejemplo: [11], [12], [13], [14], etc.

Como ya se ha señalado, la esencia del modelo *VPRSM* viene dada por la generalización que se hace del concepto estándar de relación de inclusión de conjuntos. El cual se sabe que es demasiado riguroso para representar una inclusión de conjuntos *casi* completa. A partir del concepto extendido para esta relación que se define en el modelo de *VPRSM* [15], se permite establecer o preveer un cierto grado de error en la clasificación.

Definición 1 —***Relación de inclusión estándar***: Sea U un universo finito de objetos. Sean $X, Y \subset U$, $X \neq \phi$, $Y \neq \phi$. Decimos que X está incluido en Y, o $X \subseteq Y$, si $\forall x \in X$, entonces, $x \in Y$. La Fig. 1 ilustra gráficamente esta definición.

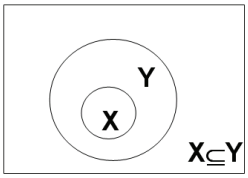


Fig. 1. Inclusión de conjuntos estándar

Resulta evidente comprender que de acuerdo a esta definición no existe la posibilidad de contemplar ningún tipo de desclasificación. Antes de establecer una definición más general para esta relación, definamos la medida del grado de desclasificación relativo del conjunto X con respecto al conjunto Y, $c(X, Y)$.

Definición 2 —***Medida del grado de desclasificación*:**

$$c(X,Y)\begin{cases}1-|X\cap Y|/|X| \\ 0\end{cases}$$

- si $X \subseteq Y \Rightarrow |X \cap Y| = |X|$, por tanto:
  $c(X, Y) = 1 - 1/1 = 0 \Rightarrow$ no hay error en la clasificación
- si $c(X, Y) \approx 1 \Rightarrow X, Y$ se acercan a ser disjuntos
- si $c(X, Y) = 1 \Rightarrow |X \cap Y| = 0 \Rightarrow X, Y$ son disjuntos

La expresión numérica *c(X, Y)* es un indicativo del error relativo de clasificación. El producto *c(X, Y)**|*X*| indicará el error absoluto de clasificación, o sea, el número de objetos mal clasificados.

Si se toma como base la medida de desclasificación relativa, se puede definir la relación de inclusión, obviando poner de forma explícita el cuantificador general, de la siguiente forma: $X \subseteq Y \Leftrightarrow c(X, Y) = 0$. Según esta definición, *c(X, Y)* puede tomar valores mayores que 0 (sin ser "demasiado grandes") en el caso que la relación sea para representar una "mayoría". O sea, se necesita que una "mayoría" de objetos de *X* sea clasificada en *Y*. Es obvio que el concepto de *mayoría* impone el establecimiento de un límite. En tal caso, se asume que *la mayoría* implica que más del 50% de los elementos de *X* deberían ser comunes con *Y*. Se añade entonces a la definición de la relación de inclusión, la especificación de un límite admisible de error en la clasificación [16].

Definición 3 —*Relación de inclusión mayoritaria*: Sea *U* un universo finito de objetos. Sea β, $0 \leq \beta < 0{,}5$, el error de desclasificación admisible. Sean $X, Y \subset U$, $X \neq \phi$, $Y \neq \phi$. Decimos que *X* está incluido mayoritariamente en *Y*, o que *X* está incluido en *Y* con un β*-error,* $X \overset{\beta}{\subseteq} Y$, si y solo si $c(X, Y) \leq \beta$. De la misma definición se puede apreciar que $\beta = 0$ expresa una *relación de inclusión estándar*, a la cual se le llama en este modelo, inclusión total**.**

En el ejemplo ilustrado en la Fig. 2 se supone que se tienen los siguientes conjuntos: $X_1 = \{x_1, x_2, x_3, x_4\}$, $X_2 = \{x_1, x_2, x_5\}$, $X_3 = \{x_1, x_6, x_7\}$, $Y = \{x_1, x_2, x_3, x_8\}$ y se ilustra la *relación de inclusión mayoritaria* entre X_1, X_2, X_3 e *Y*. Obsérvese en ella el grado de desclasificación existente entre esos conjuntos y el conjunto Y. Obsérvese además que a partir de la definición de inclusión mayoritaria dada, no se cumple $X_3 \overset{\beta}{\subseteq} Y$ pues entre esos dos conjuntos el error de desclasificación $\beta > 0{,}5$.

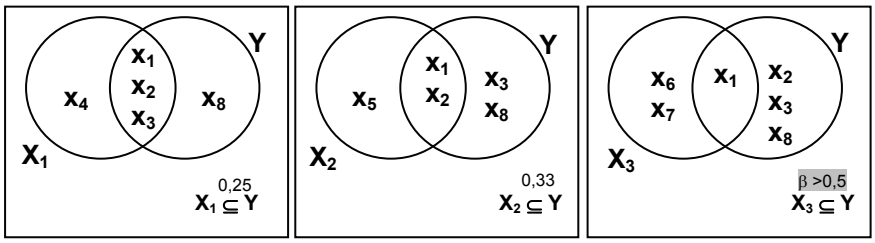


Fig. 2. Ejemplo de inclusión mayoritaria

A partir de la nueva definición de relación de inclusión, se redefinen los conceptos más representativos de *RSBM*:

Definición 4: Sea *X* un subconjunto arbitrario del universo *U*. Sea $\theta \subseteq U \times U$ una relación de equivalencia que particiona a *U* en un conjunto finito de clases de equivalencia $\langle x \rangle_\theta$. Se definen:

a) $\underline{X}_\beta = \cup \{\langle x \rangle_\theta : \langle x \rangle_\theta \overset{\beta}{\subseteq} X\}$ y se sabe que $\langle x \rangle_\theta \overset{\beta}{\subseteq} X \Leftrightarrow c(\langle x \rangle_\theta, X) \leq \beta$

b) $\overline{X_\beta} = \cup \{\langle x\rangle_\theta : \langle x\rangle_\theta \overset{\beta}{\subset} X^c\}$. Se demuestra: $\langle x\rangle_\theta \overset{\beta}{\subset} X^c \Leftrightarrow c(\langle x\rangle_\theta, X) < 1-\beta$

c) BN_β (región β-frontera) = $\overline{X_\beta}$ - $\underline{X}_\beta$

d) B^β (región β-frontera interna) = $X \cap BN_\beta$

e) NEG_β (región β-negativa) = U- $\overline{X_\beta}$

En la Fig. 3 puede apreciarse como *RSBM* planteado es un caso particular del modelo de precisión variable. En dicha figura se muestran las regiones representativas del modelo básico para un error de clasificación $\beta=0$. En tal situación, el modelo *VPRSM* se corresponde con *RSBM*.

Por su parte, en la Fig. 4 se aprecia como varían las regiones significativas, si se permite un cierto error de clasificación. En este caso, por ejemplo, se asume $\beta=0,1$. Note además que la región β-*negativa* de X es la unión de todas las clases de equivalencia que pueden clasificarse dentro de X^c, con un error en la clasificación no mayor que β.

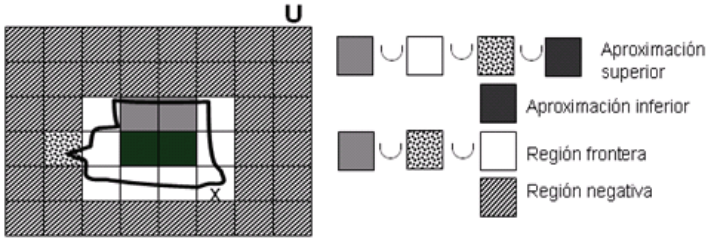


Fig. 3. Regiones representativas para $\beta=0$. Correspondencia con *RSBM*

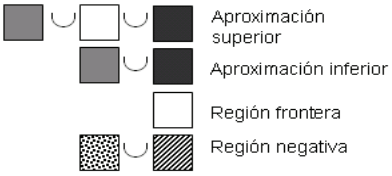


Fig. 4. Variación de las regiones significativas permitiendo un error de desclasificación $\beta=0,1$

Teniendo en cuenta que cuando β=0, el modelo de *RS* estándar es un caso particular del modelo *VPRSM*, se establece la siguiente proposición donde se expresan otras relaciones que también se cumplen.

Proposición 5:

a) $X \subseteq X_\beta$

b) $\overline{X_\beta} \subseteq \overline{X}$

c) $BN_\beta \subseteq BN$

d) $NEG \subseteq NEG_\beta$

Intuitivamente, puede observarse como al disminuir el error de clasificación β, el tamaño de la región positiva y de la región negativa de X disminuye mientras que el de la región frontera aumenta. La Fig. 5 muestra la variación de las regiones aproximadas a partir de la variación del β-*error* e ilustra y resume muchas de las propiedades que han sido enunciadas.

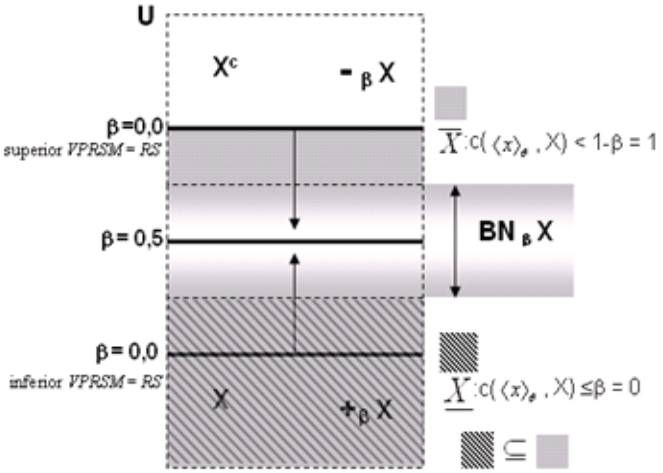


Fig. 5. Variación de las regiones significativas a partir de la variación del β-*error*

Todos estos aspectos que caracterizan esencialmente a *VPRSM* han sido tenidos en cuenta al diseñar el algoritmo de detección que se describe en la siguiente sección.

4 Algoritmo de detección de *outliers* basado en *VPRSM*

De forma sintetizada puede decirse que la principal modificación hecha a la concepción del algoritmo de detección de *outliers* basado en *RSBM* radica en el cálculo de las regiones significativas según el marco teórico propuesto en *VPRSM*. Especialmente, en lo que respecta a la determinación de las β-*fronteras internas* ($B^{\beta}{}_{i}$, $1 \leq i \leq m$). Como ya se ha señalado, en dicho modelo se permite un cierto β-*error* en la clasificación, lo cual se traduce objetivamente en flexibilizar las relaciones de inclusión a la hora de establecer las regiones significativas del modelo en el marco del análisis. Con ello, se da la posibilidad de una clasificación *casi* completa relajando de esta forma el carácter determinista de la misma según la concepción de *RSBM*.

A las entradas del algoritmo implementado para *RSBM* se le añade el β- *error*, por tanto para el algoritmo basado en *VPRSM* las entradas son las siguientes: el universo U, el concepto C, los criterios que distinguen a las relaciones de equivalencia tenidas en cuenta en el análisis (r_i, $1 \leq i \leq m$), el valor del umbral de detección μ establecido y el β-*error*. Se mantienen las mismas estructuras de datos descritas para el algoritmo basado en *RSBM*.

A continuación se presenta una versión en pseudocódigo de las dos etapas fundamentales del algoritmo: Formación de *β-fronteras internas* y proceso de detección de *outliers*.

Etapa 1 —Formación de *β-fronteras internas*: Se aplican los clasificadores (uno por cada relación) a los elementos de *U* para formar las *β-fronteras internas* para cada una de las relaciones de equivalencia tenidas en cuenta en el análisis. Complejidad temporal *O(nxmxc), c:* costo de clasificar cada elemento, *n:* cardinalidad del universo, *m:* número de relaciones de equivalencia tenidas en cuenta en el análisis.

```
(1) for r in ℜ : // para cada relación de equivalencia:
(2)             Clasificar cada elemento de U según r y definir la
                partición P_r
(3)             for clase in P_r        // por cada clase de equivalencia
                                        // definida en Pr
(4)                     if c(clase, X) ≤ β
(5)                     then
                        - por Definición 4a:
                                β
                        clase ⊆ X ⇒ clase ∈ X_β
(6)             else if c(clase, X) ≥ 1-β
(7)             then
                        -   por Definición 4b:
                                β
                          clase ⊄ X̄ ⇒ clase ∈ NEG_β
(8)                     else
                                - por Definición 4c:
                                (clase ∩ X) ⊆ B_r^β
                                // Agregar los elementos de clase que
                                // cumplen el concepto, a la frontera
                                // interna relativa a r
(9)                             B_r^β = B_r^β ∪(clase ∩ X)
```

Alg. 1. Etapa 1: Formación de las *β-fronteras internas*

```
(1) E=φ
(2) Construir las β-fronteras internas
(3) for i:= 1 to m // por cada fronterna interna
(4)             if ∀j, 1≤j≠i<m : B_j^β ⊄ B_i^β
                // si ninguna frontera interna es subconjunto de la que
                // se analiza, entonces todos los elementos de la
                // frontera interna Bi conforman el conjunto Ei
(5)             then
(6)             E = E ∪ B_i^β
(7) Outliers = {x: x∈E ^ Grado_de_excepcionalidad(x)≥ μ}
```

Alg. 2. Etapa 2*:* Proceso de detección de *outliers*

Etapa 2 —Proceso de detección de *outliers*. Se conforma el conjunto que contendrá a todos los elementos que cumplen el concepto y pueden ser candidatos a *outliers*. De

ellos, todos los que su grado de excepcionalidad sea mayor que el umbral de detección μ establecido son clasificados como tal. Complejidad temporal $O(n \times m^2)$.

Teniendo en cuenta las etapas 1 y 2, el tiempo de ejecución para todo el algoritmo es $O(máx(O(etapa\ 1), O(etapa\ 2))) = O(etapa\ 2) = O(n \times m^2)$.

En general el número de relaciones de equivalencia que intervienen en el análisis, en la inmensa mayoría de los casos, no es muy grande en relación al número de elementos del *data set*, por tal motivo, la dependencia cuadrática del tiempo de ejecución con respecto a la cantidad de relaciones de equivalencia no afecta en gran medida el tiempo de ejecución del algoritmo. Como se verá en los resultados obtenidos, esta dependencia cuadrática es casi lineal para valores pequeños ($m \leq 20$).

Con relación a la complejidad espacial, como se mantienen las estructuras de datos descritas para la versión del algoritmo basada en *RSBM*, se mantiene también el mismo orden, $O(n*m)$.

El ejemplo mostrado en la siguiente sección ilustra cómo varían las regiones significativas al permitirse un cierto β-*error* y, por tanto, cómo se flexibiliza la clasificación.

5 Detectando *outliers* en un *data set*

Se considera un universo U que representa a 25 pacientes (Tabla 1). En la tabla, por cada paciente, en función de su temperatura y a partir de la existencia o no de dolor de cabeza, se establece un diagnóstico en cuanto al padecimiento de una gripe.

Se definen dos criterios. Cada uno de los cuales particiona a U en un número determinado de clases de equivalencia.

$$r_1 = \left\{ x \in U : \begin{Bmatrix} 1\_si\_dolor\_de\_cabeza & (x) \\ 0\_en\_otro\_caso \end{Bmatrix} \right\}$$

$$r_2 = \left\{ x \in U : \begin{Bmatrix} 0\_si\_temperatura\_Normal(x) \\ 1\_si\_temperatura\_Alta(x) \\ 2\_en\_otro\_caso \end{Bmatrix} \right\}$$

CONCEPTO C = $\{x \in U \wedge gripe(x)\}$

En la Fig. 6 se muestran las clases de equivalencia que forman parte de la partición de U que se crea a partir de r_1. En ambas, hay elementos que cumplen C y elementos que no lo cumplen, por tanto, ambas clases quedan dentro de la frontera de C respecto a r_1. Los elementos de ambas clases que cumplen C son los que conforman la frontera interna. Se ilustra cómo queda la clasificación para $\beta=0$ (*RSBM*) y permitiendo un error de desclasificación $\beta=0,25$.Observe que para r_1 todo se mantiene igual en ambos casos.

Tabla 3. Datos de ejemplo que representan al universo U.

ID	Dolor Cabeza	Temperatura	Diagnóstico
1	SI	NORMAL	DESCONOCIDO
2	NO	MUY ALTA	GRIPE
3	SI	ALTA	GRIPE
4	NO	NORMAL	DESCONOCIDO
5	SI	MUY ALTA	GRIPE
6	NO	ALTA	DESCONOCIDO
7	NO	ALTA	INSOLACION
8	NO	MUY ALTA	GRIPE
9	SI	NORMAL	-
10	SI	NORMAL	INSOLACION
11	SI	MUY ALTA	GRIPE
12	NO	NORMAL	-
13	SI	NORMAL	CEFALEA
14	SI	NORMAL	CEFALEA
15	NO	MUY ALTA	GRIPE
16	NO	MUY ALTA	GRIPE
17	NO	NORMAL	-
18	NO	MUY ALTA	GRIPE
19	SI	ALTA	GRIPE
20	SI	ALTA	GRIPE
21	SI	ALTA	GRIPE
22	SI	ALTA	GRIPE
23	SI	ALTA	GRIPE
24	SI	ALTA	GRIPE
25	SI	ALTA	GRIPE

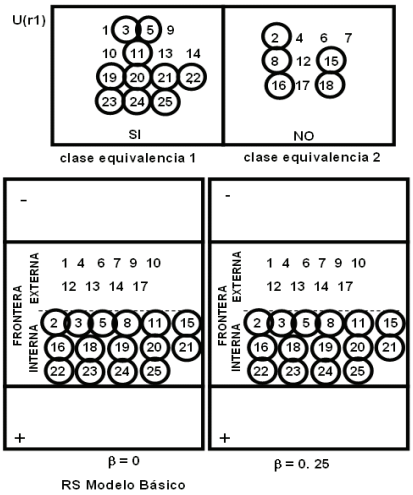


Fig. 6. Partición que establece r_1 sobre U y frontera de X respecto a r_1. $\beta=0$; $\beta=0,25$

Sin embargo, cuando se analiza lo que sucede con relación a r_2 (ilustrado en la Fig. 7), se observa que la clase de equivalencia 2 —la cual, al trabajar con el modelo básico, quedaba en la frontera y no clasificaba dentro de la región positiva aún cuando más del 80% de sus elementos cumplían con el concepto— ahora, al permitirse un error de

desclasificación $\beta=0,25$, pasa a la región positiva. Esto tiene mucho más sentido teniendo en cuenta el porcentaje de elementos de la misma que cumplen *C*.

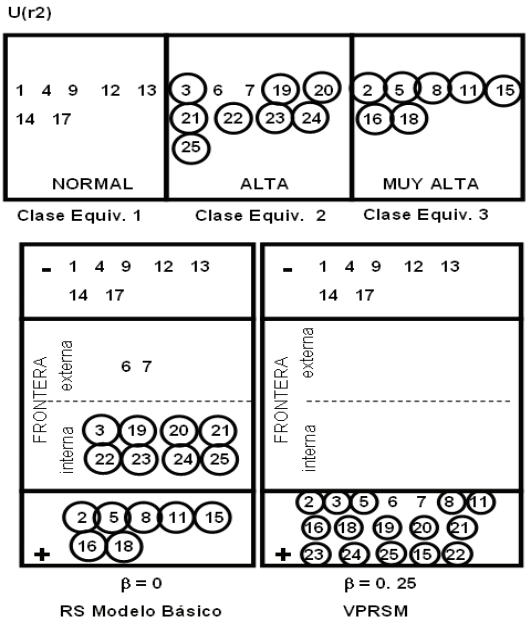


Fig. 7. Partición que establece r_2 sobre *U* y frontera de *X* respecto a r_2. $\beta=0$; $\beta=0,25$

6 Validación de los resultados. Pruebas con *data sets*

Las pruebas realizadas tuvieron como objetivo fundamental establecer comparaciones entre el algoritmo de detección basado en *RSBM* y el basado en *VPRSM* en cuanto a su complejidad temporal (a partir del análisis teórico que se hizo de este parámetro) y en cuanto a la calidad en la detección.

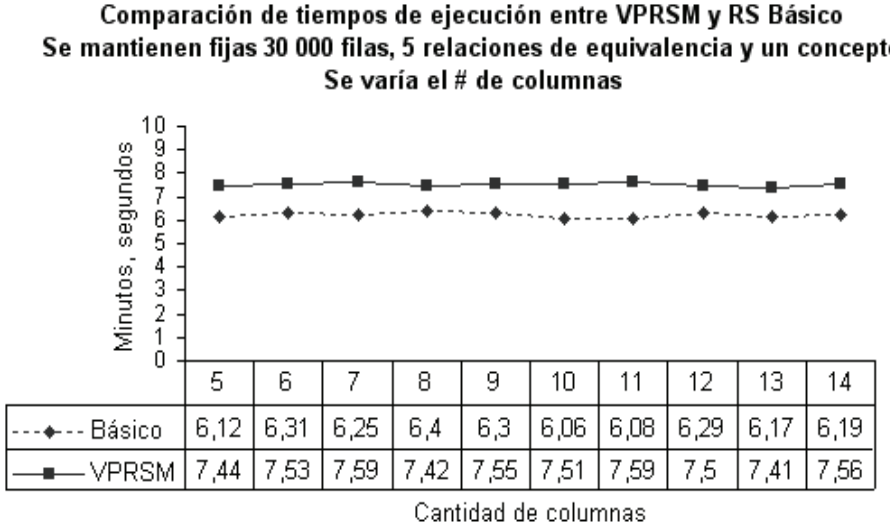
Los datos fueron obtenidos del *UCI Machine Learning Repository* del *Center for Machine Learning and Intelligent Systems* de la Universidad de California, Irvine [17]. Las pruebas, fueron hechas con un *data set* tomado de este sitio que contiene datos extraídos del *Census Bureau Database* de los EE.UU. Más de 50 artículos científicos referencian el uso de este *data set.* En [17] pueden apreciarse las características más sobresalientes del mismo y una explicación detallada de sus atributos.

Las características del equipo donde se validaron los resultados son las siguientes: PC, INTEL Pentium 4, CPU 1.5 GHz, 256 MB de RAM Plataforma: Windows XP SP3.

6.1 Tiempo de Ejecución

Las pruebas se hicieron teniendo en cuenta la variación de todos los parámetros que definen el tamaño de la entrada del algoritmo. Es decir, cantidad de filas y columnas del *data set* y número de relaciones de equivalencia que se tienen en cuenta en el análisis.

Las Fig. 8, 9, y 10 reflejan los tiempos de ejecución alcanzados por el algoritmo basado en *VPRSM* y el basado en *RSBM* en tres situaciones diferentes: Fig. 8 - se mantuvieron fijos el número de filas del *data set* (30.000), la cantidad de relaciones de equivalencia consideradas en el análisis (5) y el concepto. En este caso, se varió la cantidad de columnas del *data set*. Fig. 9 - se mantuvieron fijas las filas (30.000) y las columnas (14) del *data set*, variándose el número de relaciones de equivalencia consideradas en el análisis. Fig. 10 - la variación se hizo con respecto a la cardinalidad del *data set.*



	5	6	7	8	9	10	11	12	13	14
Básico	6,12	6,31	6,25	6,4	6,3	6,06	6,08	6,29	6,17	6,19
VPRSM	7,44	7,53	7,59	7,42	7,55	7,51	7,59	7,5	7,41	7,56

Fig. 8. *RS vs. VPRSM* en cuanto a tiempo de ejecución. Variando número de columnas

Como primera interpretación de los tres resultados, puede plantearse que se corroboran los órdenes de complejidad temporal de las dos versiones del algoritmo que fueron justificados desde el punto de vista teórico. Los resultados demuestran además, que las constantes de los órdenes son razonables y permiten que dichos algoritmos sean usados en la vida real. Otro aspecto importante a resaltar es que los tiempos de ejecución para ambas versiones del algoritmo no difieren significativamente.

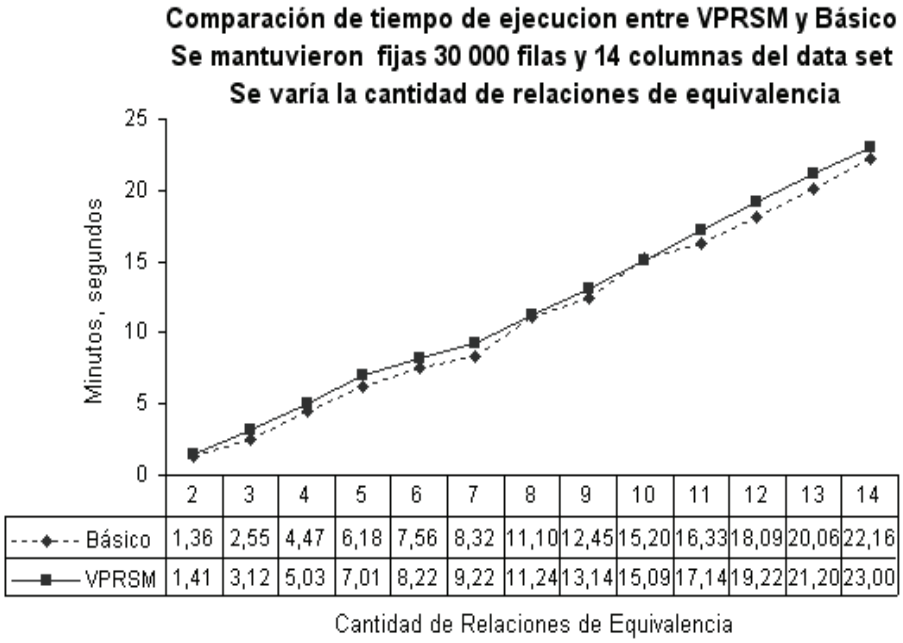


Fig. 9. *RS vs. VPRSM* en cuanto a tiempo de ejecución. Variando número de relaciones

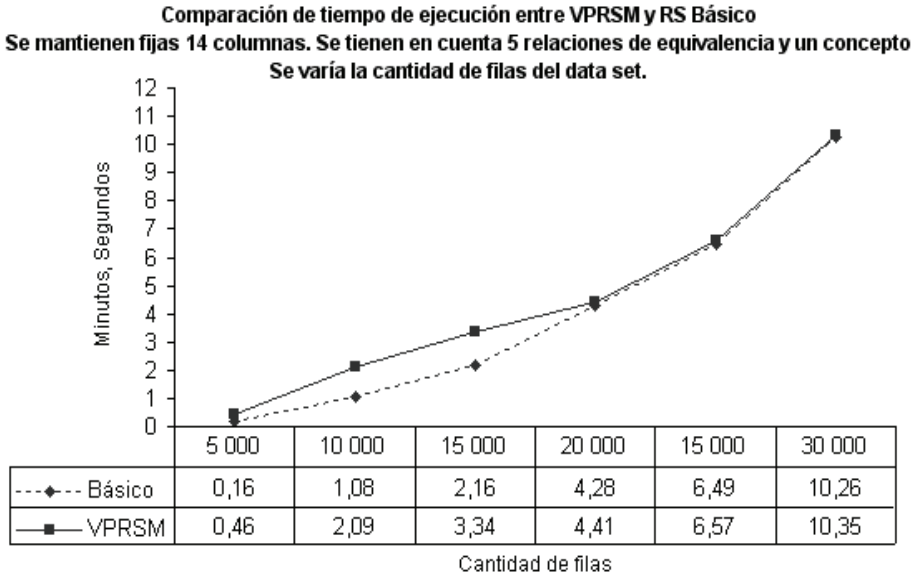


Fig. 10. *RS vs. VPRSM* en cuanto a tiempo de ejecución. Variando número de filas

6.2 Detección

En las pruebas para medir la calidad en la detección se seleccionaron los siguientes parámetros:

- Los individuos del *data set* que fueron objeto de estudio son los que cumplían con el siguiente *CONCEPTO*: 1 ≤ personas_con_edad ≤ 10.
- Los criterios a partir de los cuales se hizo el análisis quedaron establecidos por las siguientes relaciones de equivalencia:

```
r1: definida a partir del atributo categórico "workclass"
  -c1.1: workclass =['private' OR 'self-emp-not-inc' OR 'self-
      emp-inc' OR 'federal-gov local-gov' OR 'state-gov without-
      pay']
  -c1.2: workclass = ['never-worked']
r2: definida a partir del atributo categórico "education"
  -c2.1: education = ['bachelors' OR 'some-college' OR '11th' OR
      '9th' OR '7th-8th' OR '12th' OR '10th' OR 'HS-grad' OR 'prof-
      school' OR 'assoc-acdm' OR 'assoc-voc' OR 'masters' OR
      'doctorate']
  -c2.2: education = ['preschool' OR '1st-4th' OR '5th-6th']
r3: definida a partir del atributo categórico "marital-status"
  -c3.1: marital-status = ['married-civ-spouse' OR 'divorced' OR
      'separated' OR 'widowed' OR 'married-spouse-absent' OR
      'married-AF-spouse']
  -c3.2: marital-status = ['never-married']
r4: definida a partir del atributo categórico "occupation"
  -c4.1: occupation = ['tech-support' OR 'craft-repair' OR
      'other-service' OR 'sales' OR 'exec-managerial' OR 'prof-
      specialty' OR 'handlers-cleaners' OR 'machine-op-inspct' OR
      'adm-clerical' OR 'farming-fishing' OR 'transport-moving'
      OR 'priv-house-serv' OR 'protective-serv' OR 'armed-
      Forces']
  -c4.2: occupation = ['student']
```

Cualquier elemento que cumpla el concepto y pertenezca a clase c*x*.1 (*x* = 1, 2, 3, 4) es contradictorio por la relación **r***x*, pues los individuos sujetos al análisis son niños entre 1 y 10 años.

Intencionalmente, en el *data set* se introdujo un conjunto de *outliers* el cual se muestra en la Tabla 2. En ella sólo se reflejan los valores de los atributos que son relevantes para el análisis y los restantes atributos son obviados por ser irrelevantes al mismo. En esta tabla, además, puede observarse que hay valores de atributos resaltados en **negrita** e *itálica* lo cual significa que dichos valores son contradictorios para niños con edades entre 1-10. En el conjunto de *outliers* introducido el nivel de contradicción de los individuos varía. En algunos casos son contradictorios por uno o dos atributos, mientras que, en otros, lo son por tres o por cuatro y éstos son, precisamente, los elementos más contradictorios.

La gráfica que se presenta en la Fig. 11. muestra la cantidad de *outliers* detectados para diferentes valores de los umbrales β y μ. Los resultados que corresponden a *RSBM* son exactamente los alcanzados para $\beta=0$. Los valores $\beta=0{,}10;\ 0{,}20;\ 0{,}30;\ 0{,}40;\ 0{,}50;$ establecen valores de error admitidos en la clasificación y por tanto se corresponden con *VPRSM*.

Tabla 2. *Outliers* introducidos en el *data set*

Age	WorkClass	Education	Marital-Status	Occupation
7	***self-emp-inc***	1st-4th	never-married	student
6	never-worked	***masters***	never-married	student
9	never-worked	***doctorate***	never-married	student
9	never-worked	5th-6th	never-married	***Armed-Forces***
7	never-worked	1st-4th	never-married	***Adm-clerical***
8	***self-emp-inc***	***masters***	never-married	Student
8	never-worked	***doctorate***	***married-civ-spouse***	Student
6	never-worked	1st-4th	***divorced***	***Armed-Forces***
9	***federal-gov***	5th-6th	never-married	***Adm-clerical***
3	***self-emp-inc***	***masters***	***married-civ-spouse***	Student
7	never-worked	***doctorate***	***divorced***	***Adm-clerical***
2	***federal-gov***	***masters***	***divorced***	***Armed-Forces***
8	***self-emp-inc***	***doctorate***	***married-civ-spouse***	***Armed-Forces***

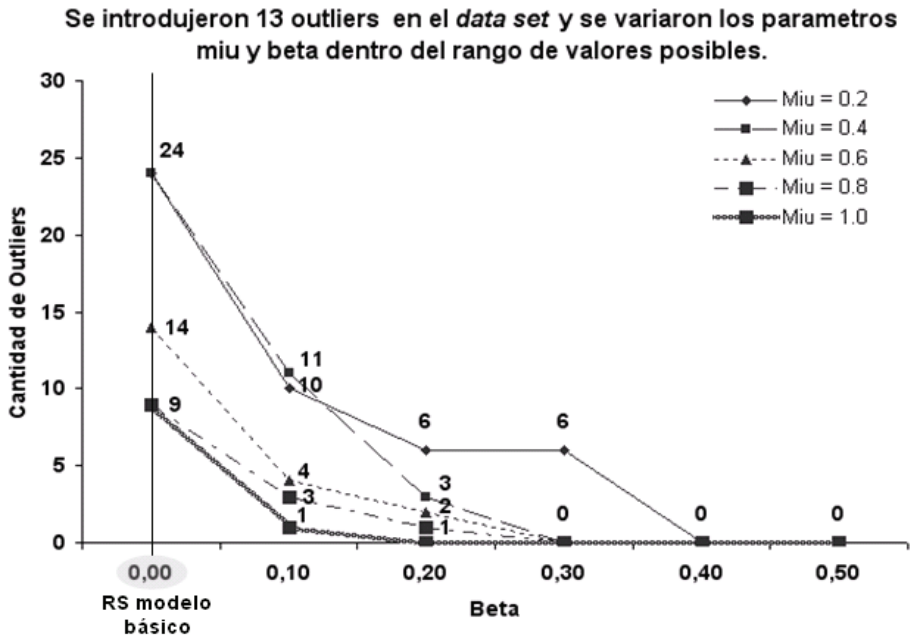


Fig. 11. *RS* modelo básico *(RSBM) vs. VPRSM* en cuanto a detección de *outliers*

El objetivo de esta prueba es mostrar la variación que experimenta la cantidad de outliers detectados a medida que se varía el valor de los umbrales β y μ. Permite,

además comparar lo que sucede en tal sentido cuando se trabaja con RSBM (β=0) y lo que sucede cuando se trabaja con VPRSM (β≠0).

Al interpretarse los resultados se debe destacar en primer lugar que, en todos los casos, dentro del conjunto de *outliers* detectados siempre se encontraron algunos de los que fueron introducidos intencionalmente en el *data set*. Cuando la cantidad de *outliers* detectados fue mayor que la cantidad de *outliers* introducidos, entonces dentro de los detectados estaban todos los introducidos. Cuando el número de *outliers* detectados fue menor que la cantidad de introducidos, entonces, de ellos, los que resultaron detectados fueron siempre los más contradictorios. Los dos ejemplos siguientes ilustran lo planteado:

- $\mu = 0.02$; $\beta = 0.0$: Se detectaron 24 *outliers*, entre ellos, estaban los 13 introducidos.
- $\mu = 0.6$; $\beta = 0.2$: Se detectaron solo 2 *outliers* que coinciden con dos de los 13 introducidos y especialmente los dos más contradictorios, pues lo eran por los cuatro atributos tomados en consideración.
- La interpretación de las pruebas realizadas nos permite, además, llegar a las siguientes conclusiones:
- Una adecuada elección de las relaciones de equivalencia o criterios de clasificación garantizan la efectividad en la detección.
- Para valores pequeños de los parámetros μ y β, en ocasiones el número de *outliers* detectados es alto y se detectan como tal elementos que realmente no lo son. Por ejemplo, para $\mu = 0.2$ y $\beta = 0.0$ se detectaron 24 *outliers*. Esto reafirma un aspecto importante de la visión estadística del problema de la detección de *outliers* para la designación final de un caso como excepcional: cuando las observaciones candidatas a ser consideradas como tal han sido identificadas por algún método de detección entonces, posteriormente a ello, el investigador debe hacer un análisis de estos resultados y seleccionar aquellas observaciones que demuestran una contradicción real con respecto a la muestra estudiada.
- Al ir aumentando sucesivamente el valor del umbral de detección (μ) se logra un refinamiento en la detección. Por lo general, una vez que el valor de dicho parámetro aumenta, disminuye el número de *outliers* detectados. Ante esta disminución, puede observarse que los que van quedando en cada caso, son los que son contradictorios por una mayor cantidad de atributos. Sin embargo, en algunos casos y para ciertas variaciones del valor de μ, no se logra un tal refinamiento. Por ejemplo, el número de *outliers* detectados al variar el valor de μ de *0,2* a *0,4,* con $\beta = 0,0$, la cantidad de *outliers* detectados, en ambos casos, es 24. Lo mismo sucede cuando se varia μ de *0,8* a *1,0,* con $\beta = 0,0$. En ambos casos también, la cantidad de *outliers* detectados fue la misma (9). Nótese que en los dos ejemplos el valor de $\beta = 0,0$ lo cual implica que no se ha permitido ningún grado de desclasificación, por tanto, son resultados referidos a *RSBM.* Nótese además que, al ir permitiendo un cierto grado de desclasificación (valores de $\beta \neq 0,0$) para las mismas variaciones de los valores de μ que las referidas en el ejemplo anterior, la cantidad de *outliers* detectados es diferente.

- Otro elemento a destacar es que, una vez que μ alcanza el mayor valor posible, $\mu=1,0$, el número de *outliers* detectados es 9, sin embargo, nuevamente hacer variaciones en el valor de β, se alcanza un mayor refinamiento en la detección, detectándose finalmente como *outliers* los elementos más contradictorios. Esto demuestra que al permitirse un grado de desclasificación controlado (β) y al variarse éste progresivamente, se mejora la calidad en la detección. A pesar de lo anteriormente dicho, se debe ser cauteloso en la variación del valor de β, pues permitir un alto grado de desclasificación puede implicar que se *limpien* completamente las fronteras internas y, por tanto, no se detecte ningún *outlier*. En las pruebas realizadas, por ejemplo, se evidencia que esto sucede a partir de $\beta = 0,3$.

7 Conclusiones

Como conclusión general puede señalarse que los resultados alcanzados a partir de las pruebas realizadas demuestran que aplicando el algoritmo basado en *VPRSM* se logra eliminar el carácter determinista en cuanto a la clasificación que limita al algoritmo basado en el modelo de *RS* básico. En tal caso, al permitirse un β-*error* (grado de desclasificación controlado) se refina la detección de *outliers* a través de una *limpieza* paulatina de las fronteras internas y con ello se logra mayor precisión en la detección haciendo que finalmente queden como *outliers* los elementos más contradictorios.

La versión del algoritmo basado en *VPRSM* además de lograr mejores resultados en la detección de *outliers* mantiene el mismo orden de complejidad temporal y espacial del algoritmo basado en *RSBM*. Aportar soluciones computacionales eficientes, tales como la posibilidad de utilizar algoritmos cuasi-lineales, es una ventaja que cualquier analista/ingeniero de datos valorará sin duda en su justa medida, dada la elevada complejidad habitual de los procedimientos en el campo del *KDD – DM*.

Como trabajo futuro, tomando como base los resultados anteriores y teniendo en cuenta que en los dos algoritmos (*RSBM* y *VPRSM*) los umbrales manejados (μ y β) deben ser definidos por el usuario, se propone como objetivo inmediato el diseño y la instrumentación de un algoritmo que *libere* al usuario, en gran medida, de tal responsabilidad. A partir del mismo y con una adecuada complejidad temporal y espacial se pretende determinar, para cada elemento de un universo finito *U*, el conjunto de valores de dichos umbrales bajo los cuales tal objeto será *outlier* en *U*. Finalmente, tomando como base el resultado anterior, se espera lograr predicciones probabilísticas sobre la posible condición de *outlier* para todo objeto del universo *U*.

Referencias

1. Pawlak, Z., (1982) Rough Sets. International Journal of Computer and Information Sciencies, 11(5):341-356.
2. Jiang, F., Sui, Y., & Cao, C., (2005). Outlier detection using rough sets theory. Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing (RSFDGrC 2005). Springer.

3. Ziarko, W., (1993). Variable Precision Rough Set Model. s.l. : Journal of Computer and System Sciences, 46(1):39-59.
4. Ching-Hsue, C., You-Shyang, C., & Jr-Shian, C. (2007). Classifying Initial Returns of Electronic Firm's IPOs Using Entropy Based Rough Sets in Taiwan Trading Systems. ICICIC'07, The Second International Conference on Innovative Computing, Information and Control. Kumamoto, Japan: IEEE Computer Society.
5. Sang Wook, H., & Jae-Yearn, K. (2007). Rough Set-based Decision Tree using the Core Attributes Concept. ICICIC'07, The Second International Conference on Innovative Computing, Information and Control. Kumamoto, Japan: IEEE Computer Society.
6. Beynon, M. J. (2006). An introduction of the condition class space with continuous value discretization and rough set theory. Wiley InterScience; International Journal of Intelligent Systems , Volume 21, Issue 2, Pages: 173-191.
7. Hirokane, M., Konishi, H., Miyamoto, A., Nishimura, F. (2007). Extraction of minimal decision algorithm using rough sets and genetic algorithm. Wiley InterSciences, Systems and Computers in Japan , Volume 38, Issue 4, Pages: 39-51.
8. Strömbergsson, H., Prusis, P., Midelfart, H., Lapinsh, M., Wikberg, J., & Komorowski, J. (2006). Rough set-based proteochemometrics modeling of G-protein-coupled receptor-ligand interactions. Wiley InterScience. Proteins: Structure, Function, and Bioinformatics , Volume 63, Issue 1, Pages: 24-34.
9. Pawlak, Z., (1991). Rough Sets: Theoretical Aspects of Reasoning About Data. s.l.: Spinger.
10. Ziarko, W., (2001). Probabilistic Decision tables in the Variable Precision Rough Set Model. Computer Science Department, University of Regina, Regina, Saskatchewan, S4S 0A2, Canada.
11. Gong, ZT., Sun, BZ., Shao, YB., Chen, DG., He, Q. (2004) Variable precision rough set model based on general relations. Proceedings of the Third Conference on Machine Learning and Cybernetics, Shanghai.
12. Beynon, M. J., Driffield, N. (2005). An illustration of variable precision rough sets model: an analysis of the findings of the UK Monopolies and Mergers Commission. Computers and Operations Research Volume 32, Issue 7, Pages: 1739-1759.
13. Su, ChT., Hsu, JH. (2006)**.** Precision parameter in the variable precision rough sets model: an application**.** Volume 34, Issue 2, Pages: 149-157.
14. Maheswari, V.U., Siromoney, A. Mehata, K.M. (2001). The Variable Precision Rough Set Model for Web Usage Mining. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, Volume 2198/2001. Web Intelligence Research and Development, pages: 520-524
15. Ziarko, W., (1994). Rough Sets, Fuzzy Sets and Knowledge Discovery, Springer Verlag, 326-334.
16. Ziarko, W., (1999). Decision making with probabilistic decision tables. Proceedingas of the 7th International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing (RSFDGrC' 99). Yamaguchi, Japan, Lectures Notes in AI 1711, Springer Verlag, pp. 463-471.
17. *UCI Machine Learning Repository*. http://cml.ics.uci.edu. Última consulta: 30/05/09