Algoritmo para la detección de casos excepcionales basado en la Teoría de Conjuntos Aproximados

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, España
pmacia@dtic.ua.es

Resumen. Los *outliers* son objetos que muestran un comportamiento *anormal* dentro del contexto donde se encuentran o que tienen valores inesperados en algunos de sus parámetros. Por la importancia que ello reviste en los procesos de búsqueda de información en grandes volúmenes de información, los investigadores prestan especial atención al desarrollo de técnicas de detección eficientes. Este artículo tiene como antecedentes una investigación previa que basa el problema de la detección de *outliers* en la teoría de *Rough Sets*. En dicho trabajo, los *outliers* se definen como elementos de los *conjuntos excepcionales no redundantes* que poseen un *grado de excepcionalidad* mayor que un umbral establecido. En el presente artículo se hace una revisión crítica de este trabajo evidenciando que resulta impracticable una implementación computacional de un algoritmo para detectar *outliers* a partir de dicho planteamiento teórico por ser de orden exponencial y se propone una extensión del marco teórico original sobre el que se diseña un algoritmo de complejidad temporal no exponencial basado en el método de detección propuesto junto con una implementación que ha permitido validar la propuesta.

1 Introducción

Las investigaciones más recientes en aspectos relacionados con el *Knowledge Discovery in Databases (KDD)* prestan especial atención a la detección de casos excepcionales (*outliers*) que puedan observarse en los grandes volúmenes de información que almacenan las extensas bases de datos de hoy día. Si, en general, los procesos de *KDD - Data Mining* se enfocan en el sentido de descubrir patrones de comportamiento representativos (alta fiabilidad y soporte), la detección de *outliers* aprovecha justamente la elevada marginalidad de estos objetos para detectarlos midiendo su grado de desviación respecto a dichos patrones.

En ausencia de una definición formal, la de D.Hawkin [1] describe acertadamente un *outlier* por su comportamiento: «un *outlier* es una observación que se desvía tanto de las otras, que permite suponer que se ha generado por un mecanismo diferente». Barnett y Lewis [2], por su parte, proponen otra definición donde se aprecia una esencia similar: «un *outlier* en un conjunto de datos, no es más que una observación, o un conjunto de observaciones, que parece ser inconsistente con el resto de los datos».

Una eficiente y oportuna detección de *outliers* puede evitar que se corra el riesgo de tomar malas decisiones basadas en datos erróneos, además de ayudar en la detección, prevención y reparación de los efectos negativos que puede traer consigo el uso indebido de los mismos. La presencia de *outliers* en un conjunto de datos puede entorpecer notoriamente la detección de patrones confiables dentro del proceso de minería de datos, donde este aspecto es uno de los objetivos fundamentales del mismo. En otros casos, puede que su detección, como casos distinguidos por su excepcionalidad, sea el objetivo fundamental del propio proceso de minería de datos.

El problema de la detección de *outliers* puede tener incidencia directa en múltiples contextos de aplicación. A continuación se mencionan algunos de ellos: detección de intrusos en las redes [3], diagnóstico médico [4], aspectos socio culturales [5], estudios climáticos [6], *e-business* [7], en el contexto del *data Warehouse* [8], video-vigilancia (*video surveillance*) [9], análisis de datos químicos [10], etc.

Todo este conjunto de aplicaciones de actualidad en las que el problema de la detección de *outliers* tiene una incidencia directa hace evidente la existencia del mismo así como la necesidad y el interés de la comunidad científica de resolverlo con técnicas y métodos cada vez más eficientes.

Los resultados que se exponen en el presente trabajo toman como punto de partida una comunicación presentada en el congreso sobre *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing* (RSFDGrC 2005) por F. Jiang, Y. Sui, y C. Cao [11] donde se presenta un marco teórico sobre el cual se propone un método de detección de *outliers* basado en el modelo básico de la Teoría de Conjuntos Aproximados *(Rough Sets —RS)*. Dicha propuesta es de orden exponencial, haciendo impracticable una implementación computacional de un algoritmo de detección a partir de la misma. En este trabajo se propone una extensión del marco teórico original sobre el que se construye un algoritmo de complejidad temporal no exponencial. Para ello, el resto del artículo se ha organizado de la siguiente forma: en el siguiente apartado se destacan aspectos relevantes de la Teoría de *RS* y cómo su aplicación en múltiples contextos investigativos pone de manifiesto su efectividad; en el apartado 3 se realiza un análisis crítico de la propuesta teórica de Jiang et al. y se destacan las principales fortalezas y debilidades de la misma, valorando positivamente las principales fortalezas de la propuesta e identificando como un problema esencial el hecho de que una instrumentación computacional del método de detección propuesto tendría una complejidad temporal exponencial a partir de la definición de *outlier* dada; en el apartado 4 se presenta un conjunto de resultados matemáticos (lemas, corolarios y propiedades) que permiten ampliar el marco teórico propuesto y posibilitan el diseño e instrumentación computacional del algoritmo de detección de *outliers,* con orden de complejidad temporal y espacial no exponencial, presentado en el apartado 5; en el apartado 6 se dan algunos detalles sobre la implementación y; en el apartado 7 se ilustra, mediante un ejemplo, la funcionalidad del algoritmo; el apartado 8 refleja aspectos relativos a la validación de los resultados con *data sets* de

la vida real y; finalmente en el apartado 9 se presentan las principales conclusiones que se desprenden de esta investigación y se plantean las líneas futuras a seguir.

2 Antecedentes

En el decursar del tiempo han existido diversas aproximaciones al problema de la detección de *outliers* en grandes volúmenes de información. En el campo de la Estadística fue donde primero se comenzó a tratar el problema.

Los modelos estadísticos generalmente son apropiados para el procesamiento de conjuntos de datos con valores reales continuos, cuantitativos o, al menos, datos cualitativos con valores ordinales. En la actualidad, cada vez más se necesita procesar datos expresados de manera categórica (no ordinales). Esto limita considerablemente la aplicación de los métodos estadísticos. Otra deficiencia que se señala a los métodos es su limitación para trabajar correctamente en espacios de alta dimensionalidad (multivariados), donde resulta generalmente muy difícil encontrar el modelo adecuado [12].

Existen métodos basados en aproximaciones no paramétricas entre los que se encuentran los métodos de detección de *outliers* basados en distancias. Entre ellos, el más citado es el de *k-vecinos más cercanos* (*K-NN*) [13] [14]. Existen diferentes enfoques del algoritmo *K-NN* pero todos usan una métrica adecuada para el cálculo de las distancias entre *vecinos* tal como la distancia Euclidiana o la distancia de Mahalanobis. Se registran varias optimizaciones del algoritmo básico de *K-NN* [15].

En general, puede decirse que existe un numeroso conjunto de técnicas para detectar *outliers* donde se mezclan algoritmos de diversos tipos [11] [16]. Entre los métodos mas sobresalientes pueden señalarse los basados en distribuciones [17], en profundidades [18], en distancias [13] [19], en densidades [20], en *clusters* [21], en *support vector* [22], etc. Las investigaciones más recientes recogen varios métodos de detección basados en técnicas de Inteligencia Artificial, fundamentalmente, las referidas al *Machine Learning* [23] y, dentro de ellas, pueden citarse árboles de decisión y redes neuronales [11], fundamentalmente.

Una guía bastante completa de los métodos de detección de *outliers* más sobresalientes que han sido publicados se pude encontrar en [23].

Considerando que no se cuenta con una aproximación universalmente aplicable de detección de *outliers* y que los investigadores deben centrar sus esfuerzos en la selección de un método aceptable para su conjunto de datos, este tema representa aun un problema muy abierto y, en consecuencia, siguen apareciendo referencias a nuevos modelos y nuevos métodos basados en diversos enfoques y aproximaciones al problema en cuestión.

Como ya se ha expresado con anterioridad, el punto de partida de esta investigación es una comunicación presentada en la edición del 2005 del congreso *RSFDGrC (Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing)* bajo el título *Outlier detection using Rough Sets Theory*. Los estudios previos y los resultados de F.Jiang, Y.Sui y C.Cao, además de sentar las bases del presente trabajo, constituyen, a su vez, el primer antecedente de la utilización de la Teoría de *Rough Sets* en el campo de *outlier detection* [11] [24].

Tomando en consideración que el método de detección propuesto desde el punto de vista teórico-matemático se basa en la Teoría de *Rough Sets,* resulta necesario exponer los aspectos esenciales que caracterizan a dicha Teoría.

La Teoría de Conjuntos Aproximados o *Rough Sets*, propuesta en 1986 por el recientemente fallecido Profesor Z. Pawlak, de la Universidad Tecnológica de Varsovia [25], es una extensión de la Teoría de Conjuntos para su aplicación al caso de información incompleta o insuficiente. Esta teoría surge a partir de la necesidad práctica de resolver problemas de clasificación y en ella se asume que junto a cualquier objeto del universo hay asociada una cierta cantidad de información**:** el conocimiento que se tiene acerca de dicho objeto y el cual se expresa mediante valores asociados a un conjunto de propiedades que describen a dicho objeto. Esta teoría tiene el atractivo añadido de contar con una base matemática simple y sólida: la teoría de relaciones de equivalencia, que aquí permite describir particiones constituidas por clases indiscernibles que agrupan a objetos con atributos similares. Se trata de una metodología de clasificación de datos.

La aplicación de la Teoría de *Rough Sets* en múltiples contextos investigativos pone de manifiesto su efectividad y su versatilidad en la solución de disímiles problemas. Especialmente ha sido aplicada con resultados elocuentes en los procesos de *KDD - Data Mining.* Algunos ejemplos que ilustran lo anteriormente dicho son los siguientes: en el ámbito de los sistemas de mercado (*Trading Systems*) la teoría ha sido usada para fines predictivos [26], en el contexto del *Machine Learning* en la concepción de algoritmos de clasificación basados en árboles de decisión [27], en investigaciones básicas orientadas al desarrollo de Sistemas Inteligentes [28], en problemas de clasificación usando árboles de decisión [29], en el campo de la Bioinformática [30], etc.

Tenido en cuenta la simplicidad del planteamiento formal de Jiang et al. y lo novedoso de su enfoque, basado en una teoría de bases matemáticas simples y sólidas, la Teoría de *Rough Sets*, así como la capacidad de la misma para modelar un amplio espectro de situaciones reales, se valora como positivo el estudio de esta propuesta.

3 Análisis crítico sobre una propuesta de un método de detección de outliers basado en la Teoría de Rough Sets

Con el objetivo de ubicar al lector en los elementos fundamentales de la Teoría de *Rough Sets* a continuación se presentan los conceptos fundamentales de la misma que son usados en el método de detección propuesto. En [25] se puede encontrar una explicación completa de los fundamentos matemáticos en los que se basa dicha Teoría.

Sean $U \neq \phi$ el universo (finito) y $r \subseteq UXU$, una relación de equivalencia definida sobre *U*.

Sea $X \subseteq U$, un concepto. Se definen dos aproximaciones que caracterizan a *X*:

Aproximación superior: $\bar{r}(X) = \cup \{Y \in U / r : Y \cap X \neq \phi\}$. La unión de todas las clases de equivalencia inducidas por *r* en *U* cuya intersección con *X* es no vacía.

Aproximación inferior: $\underline{r}(X) = \cup \{Y \in U / r : Y \subseteq X\}$. La unión de todas las clases de equivalencia inducidas por *r* en *U* que están contenidas en *X*.

La propia teoría de *RS*, define el concepto de *frontera* de la siguiente forma

Frontera: $BN(X) = \bar{r}(X) - \underline{r}(X)$

La caracterización matemática de *outliers* propuesta por Jiang et. al. se basa esencialmente en los elementos que acabamos de enunciar y en especial en el concepto de *Frontera interna* que ellos definen.

Definición 1 —*Frontera interna:* Sea $\Re = \{r_1, r_2, ..., r_m\}$, *m* relaciones de equivalencia definidas sobre el universo U. La *frontera interna* de X con respecto a r_i, se define de la siguiente manera: $B_i(X) = BN_i(X) \cap X = X - r_i(X)$

Definición 2 —*Conjunto excepcional:* Sea $e \subseteq X$ tal que: $\forall r_i \in \Re$, $B_i(X) \neq \phi$ se cumple $e \cap B_i(X) \neq \phi$. Entonces al conjunto *e* se le llama *conjunto excepcional* de X con respecto a $\Re$.

Definición 3 —***Elemento dispensable:*** Sea e un *conjunto excepcional* de *X* con respecto a $\Re$ y sea $x \in e$ tal que *e-{x}* es también *excepcional* con respecto a $\Re$. Decimos entonces que *x* es *dispensable* en e con respecto a $\Re$.

En caso contrario, *x* es *indispensable* en *e* con respecto a $\Re$.

Definición 4 —*Conjunto excepcional no redundante:* Se dice que un *conjunto excepcional* es *no redundante* si todos sus elementos son indispensables.

Definición 5 —*Grado de Marginalidad*: Sea *x* un elemento arbitrario de *X*. El *grado de marginalidad* de *x* con respecto a $\Re$ es la cantidad de fronteras internas diferentes, de *X* con respecto a $\Re$, que contienen a *x*:

$BD(X) = \left|\{B_i(X), i = 1,2,...,m : x \in B_i(X)\}\right|$.

Definición 6 —***Grado de excepcionalidad:*** El *grado de excepcionalidad* de *x* se define de la siguiente forma: OD(x)=BD(x) / $|\Re|$

Definición 7 —*Outlier:* Un *outlier* en *X* con respecto a $\Re$ es un objeto x que pertenece a algún conjunto excepcional no redundante de *X* con respecto a $\Re$ que tiene un grado de excepcionalidad mayor que un *umbral* μ dado.

Sobre la propuesta teórica de Jiang et al. se llega a las conclusiones que se detallan en los párrafos siguientes.

- Los resultados de Jiang et al. aportan un marco teórico pero sin materializar una solución, lo cual permite desarrollar la idea.
- El enfoque propuesto es original y novedoso pues no existen, al menos en la bibliografía revisada, antecedentes de otro con un planteamiento similar. No cae dentro de ninguna de las categorías que la bibliografía recoge para clasificar los métodos de detección según el principio en el cual se basan los mismos.
- El método de detección propuesto es simple en cuanto al planteamiento teórico en que se basa, pero su implementación computacional a partir de la definición de *outlier* dada cae en un problema no tratable:

  > *«Se sabe que dado un conjunto C, con |C|= n, la cantidad total de posibles subconjuntos de C (El Conjunto Potencia de C) es 2^n. Un algoritmo de detección según la definición de outliers dada, tendría que hallar siempre el Conjunto Potencia de X para posteriormente, a partir del mismo, seleccionar los conjuntos excepcionales no redundantes, de los cuales saldrían finalmente los outliers. Por lo antes expresado, la complejidad temporal de dicho algoritmo sería $\Omega(2^n)$.»*

- El modelo Rough Set ha sido aplicado exitosamente en la solución de un gran número de problemas, lo cual demuestra su efectividad y su versatilidad. En consecuencia, el método, por estar basado en dicho modelo, se espera que también sea potente y eficaz al ser usado dentro de contextos de minería de datos en los que sea factible su aplicación pero, como el procedimiento de detección propuesto se basa en el Modelo Básico de RS [25], hay que prestar especial atención a las limitaciones que tiene el mismo: incapacidad para modelar información incierta. La clasificación con un grado controlado de incertidumbre o un posible error de clasificación, está fuera del alcance de este modelo. Sin embargo, 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. La definición estándar de inclusión de conjunto tenida en cuenta en el Modelo Básico de RS es demasiado rigurosa para modelar una inclusión de conjuntos cuasi completa. Por tanto, estas limitaciones le restan soporte al resultado.
- Uno de los elementos que puede limitar la efectividad de un método de detección de *outliers* es la naturaleza de los datos sobre los cuales se aplica el mismo. El método propuesto es aplicable tanto a datos continuos como discretos (categóricos).
- En la propuesta del método no se avizora ningún elemento que pueda limitar su aplicación en *data sets* de gran dimensionalidad, sin embargo, esto es un aspecto que limita la aplicación efectiva de varios métodos de detección recogidos en la bibliografía revisada.

Los aspectos que se acaban de señalar constituyen, a nuestro juicio, las principales fortalezas y debilidades de la propuesta teórica de Jiang. et al. Por tanto, valorando positivamente las principales fortalezas de la propuesta e identificando como un problema esencial el hecho de que una instrumentación computacional del método de detección propuesto tendría una complejidad temporal exponencial a partir de la

definición de *outlier* dada, consideramos como hipótesis de partida de nuestra investigación asumir que la limitación dada puede ser resuelta ampliando el marco teórico propuesto con nuevos resultados desde el punto de vista matemático (lemas, corolarios y propiedades) que posibiliten el diseño e instrumentación computacional de un algoritmo de detección de *outliers* basado en la propuesta de Jiang et al. con orden de complejidad temporal y espacial no exponencial.

Por lo tanto, con el objetivo de poder concebir (diseño e implementación) un algoritmo computacionalmente eficiente para la detección de *outliers* a partir de la **definición 7,** se hace necesario hacer una extensión de los resultados teóricos propuestos por Jiang et al.

4 Propuesta de ampliación del marco teórico

Los siguientes lemas y proposiciones constituyen el fundamento matemático del algoritmo de detección de *outliers* que se propondrá.

Proposición 8: $\forall e, e' \subseteq X$, si e es excepcional y $e \subseteq e'$, entonces e' es excepcional también.

Lema 9: Sea *f* un *conjunto excepcional* no redundante y sea *a*: $a \in X \wedge a \in f$, $\exists f \Leftrightarrow \exists i, 1 \leq i \leq m$, tal que $B_i^c \cup \{a\}$ es un *conjunto excepcional*.

Lema 10: Si $\exists a \in B_i, 1 \leq i \leq m$, tal que $B_i^c \cup \{a\}$ no es un *conjunto excepcional,* entonces, $\exists j, 1 \leq j \leq m, j \neq i$, tal que $B_j \subseteq B_i$.

Corolario 11: Si $\forall j, 1 \leq i, j \leq m, j \neq i$, se cumple $B_j \not\subset B_i$ (lo cual quiere decir que la *frontera interna* B_i no contiene completamente a ninguna otra *frontera interna*), entonces, $\forall a \in B_i$, $B_i^c \cup \{a\}$ es un *conjunto excepcional*.

Lema 12: Sea $a \in$ X. Si $\exists j, j \neq i, 1 \leq i, j \leq m$, tal que $B_j \subseteq B_i$ y $B_i^c \cup \{a\}$ es un *conjunto excepcional*, entonces, $B_j^c \cup \{a\}$ es también un *conjunto excepcional*.

Definición 13: Sea *f* cualquier *conjunto excepcional* no redundante. Se define el conjunto E_i, $1 \leq i \leq m$, de la siguiente forma: $E_i = \{a: a \in X, a \in f, f \cap B_i = \{a\}\}$. E_i contendrá a todos los elementos de *X* que pertenecen a algún *conjunto excepcional* no redundante (tenidos en cuenta todos ellos) y que además son miembros de la *frontera interna* B_i.

A partir de lo anterior, se llega a la siguiente conclusión: $E = \bigcup_{i=1}^{m} E_i$, es el conjunto de todos los elementos de *X* que pertenecen a algún *conjunto excepcional* no redundante.

Lema 14: $\forall i,\ 1 \leq i \leq m,$ se cumple que $E_i \subseteq B_i$, o sea, todos los elementos de algún E_i particular, son elementos de la *frontera interna* B_i.

Lema 15: Sea $a \in X,\ 1 \leq i \leq m,\ B_i^c \cup \{a\}$ es un *conjunto excepcional* si y solo si $a \in E_i$.

Todos estos resultados han sido demostrados formalmente y las demostraciones no se exponen en el presente trabajo para no extender demasiado el mismo.

En el siguiente acápite se dan detalles sobre el algoritmo propuesto y los aspectos del marco teórico sobre los cuales se basa el mismo.

5 Un algoritmo para la detección de outliers basado en el modelo básico de Rough Sets

Antes de exponer el algoritmo de detección propiamente dicho, veamos algunas consideraciones preliminares.

Sea C el *concepto* a tener en cuenta, y sea $\Re = \{r_1, r_2, ..., r_m\}$ un conjunto de relaciones de equivalencia (*criterios*) definidas sobre U.

Se calculan las fronteras internas (B_i) de C con respecto a cada elemento de $\Re$. Para cada elemento que cumple C y además pertenece a cualquier frontera interna B_i se calcula su *grado de marginalidad*. Una vez concluido esto se comienza a conformar el conjunto E a partir de los elementos de los conjuntos E_i.

Si en el análisis de un E_i particular, se detecta algún elemento de C que con anterioridad ya había sido identificado como miembro de otro E_j y, por tanto, ya había sido incluido en E, éste no se tiene en consideración.

Como suposición teórica, el algoritmo sólo tiene en cuenta fronteras internas diferentes y no vacías. Esta consideración se hace sin pérdida de generalidad, pues si hay dos fronteras iguales, los elementos de una que pertenecen a algún conjunto excepcional no redundante serían los mismos en ambos casos y, por tanto, tenerlas duplicadas no aporta nada relevante en relación a los resultados que se obtienen. Las fronteras internas vacías no son tenidas en cuenta por definición (definición 2 — Conjunto Excepcional)

La esencia del algoritmo es determinar la relación de inclusión entre fronteras internas y, en función de las mismas, tomar diferentes decisiones y realizar acciones que no son más que la aplicación directa de alguno de los Lemas y Corolarios que fueron enunciados, demostrados y que sirven de marco teórico a la concepción del algoritmo que se presenta. Finalmente, se obtiene el conjunto E, donde quedaran todos los elementos del universo U que pertenecerán a algún conjunto excepcional no redundante.

A continuación se presenta una versión en pseudocódigo del algoritmo de detección propuesto.

```
(1) for i:= 1 to m // iterar por todas las fronteras internas
(2)           if ∀j, 1≤j≠i<m : Bj ⊄ Bi
(3)           then
```

```
                - por corolario 11:
                  ∀a∈Bi, Bi^C ∪{a}es un conjunto excepcional
                - por Lema 15: a∈Ei ⇒ Ei⊇Bi
                - por Lema 14: Ei⊆ Bi
              y por tanto se concluye que Ei=Bi
// si ninguna frontera interna es subconjunto de la
// que se analiza en un momento dado, entonces
// todos los elementos de la frontera interna Bi
// conforman el conjunto Ei
(4)         else ∃j,1≤j<i,tal que Bj ⊂ Bi
// en otro caso, existe al menos una frontera interna Bj
// que es subconjunto de Bi
(5)                 if Bi^C ∪{a} es un conjunto excepcional
// si Bj⊂Bi y ∃a∈Bi tal que Bi^C ∪{a} es un conjunto
// excepcional, entonces todo elemento indispensable de
// Bi, lo será también de Bj
(6)                 then
                    - por Lema 15: a∈Ei (I)
                - por Lema 12:
                                Bj^C ∪{a} es un conjunto excepcional
                    - por Lema 15: a∈Ej (II)
                    Luego por (I) y (II): Ei⊆Ej
                    Por tanto, no es necesario construir en este
                    momento el conjunto Ei.
(7)                 else Bj^C ∪{a} no es un conjunto
                       excepcional y por Lema 15: se
                       asegura que a∉Ei
(8) Una vez que  el conjunto E ha sido construido, se detectan
los elementos del mismo cuyo grado de excepcionalidad es mayor
que el umbral previamente establecido. Dichos elementos, son los
outliers.
```

Alg. 1. Algoritmo de formación del conjunto *E*

De inmediato se dan detalles de la implementación computacional del algoritmo descrito.

6 Consideraciones sobre la implementación computacional

Los parámetros de entrada del algoritmo son los siguientes: el universo *U*, el concepto *C*, las relaciones r_i , $1 \leq i \leq m$ - $\Re = \{r_1, r_2, ..., r_m\}$ y el grado de excepcionalidad (valor del *umbral* μ) establecido.

La estructura de datos fundamental que se utiliza en el algoritmo es la de *diccionario*, entendiendo por esto un conjunto de pares (*clave*, *valor*), donde *clave* es un objeto cualquiera al cual se le asocia uno y sólo un objeto de tipo *valor*.

En el algoritmo, las *claves* se obtienen como resultado de aplicar un clasificador a un elemento cualquiera del universo. Dicho clasificador está asociado a una relación de equivalencia r_i particular, $1 \leq i \leq m$ y permite clasificar a los miembros de las clases de equivalencia definidas por dicha relación. Los *valores* asociados a las *claves* son listas de elementos que pertenecen a la clase de equivalencia identificada por la *clave* asociada a dicho *valor*.

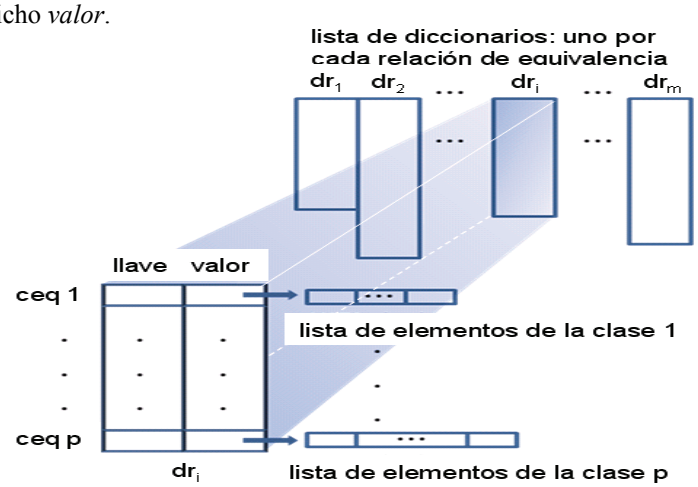

Fig. 1. Estructuras de datos utilizadas

En general, para cada relación de equivalencia se conforma un *diccionario* y a partir de todos ellos se conforma una *lista* de dimensión *m*, donde *m* es el número de relaciones de equivalencia tenidas en consideración. La Fig. 1 ilustra lo expresado.

Las entradas del algoritmo se almacenan en *listas*. La *lista* es un tipo básico, por lo que no es necesario dar detalles con respecto a su funcionalidad.

De acuerdo a las estructuras de datos utilizadas, puede decirse que la complejidad espacial del algoritmo es $O(n \times m)$ pues cada *diccionario*, a lo sumo, puede contener a todos los elementos del universo. La memoria que ocupan las restantes estructuras de datos no rebasa este orden. Donde, *n:* cardinalidad del universo y *m*: número de relaciones de equivalencia tenidas en cuenta en el análisis.

La implementación computacional del algoritmo consta de dos etapas fundamentales:

Etapa 1 —Formación de fronteras internas: se aplican los clasificadores (uno por cada relación de equivalencia) a los elementos del universo (*data set*) con el objetivo de formar las fronteras internas para cada una de las relaciones de equivalencia tenidas en cuenta en el análisis. Complejidad temporal $O(n \times m \times c)$, *c:* costo de clasificar cada elemento.

Etapa 2 —Proceso de detección de *outliers*: es una implementación concreta del algoritmo propuesto. 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 filas de la tabla. Por tal motivo, la dependencia cuadrática del tiempo de ejecución con

respecto a la cantidad de relaciones de equivalencia no afecta en gran medida al tiempo de ejecución del algoritmo. Como se verá en los resultados obtenidos, esta dependencia cuadrática es casi lineal para valores pequeños de m ($m \leq 20$).

La forma de calcular el grado de excepcionalidad de un elemento del universo teniendo previamente calculadas las fronteras internas es relativamente sencilla. Consiste en determinar, para cada elemento del universo, la cantidad de fronteras internas a las que pertenece y hallar el cociente entre dicho número y el total de fronteras internas que existen.

El ejemplo mostrado en el siguiente apartado ilustra la funcionalidad del algoritmo descrito.

7 Encontrando outliers en un data set

Se considera un universo U que representa a 21 pacientes (Tabla 1). En la base de datos, 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*** (relaciones de equivalencia tenidas en cuenta en el análisis). 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∈U ^ gripe(x)}*

En la Fig. 2. se muestran las clases de equivalencia que forman parte de la partición de U que se crea a partir de *r1*. En ambas, hay elementos que cumplen el concepto y elementos que no lo cumplen, por tanto, ambas clases quedan dentro de la frontera de C respecto a *r1*. Los elementos de ambas clases que cumplen el concepto son los que conforman la frontera interna.

Por su parte, en la Fig. 3. se muestran las clases de equivalencia que forman parte de la partición de U que se crea a partir de r2. En este caso, la clase 1 no tiene elementos que cumplen el concepto, por tanto, dicha clase es irrelevante para el análisis. La clase 2 tiene elementos que cumplen el concepto y elementos que no lo

cumplen. Los que lo cumplen pertenecen a la frontera interna. Por su parte, todos los elementos de la clase 3 cumplen el concepto y por ello dicha clase está incluida completamente en la aproximación inferior de r2.

Tabla 2. 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	insolación
8	no	muy alta	gripe
9	si	normal	-
10	si	normal	insolación
11	si	muy alta	gripe
12	no	normal	-
13	si	normal	cefalea
14	si	normal	cefalea
15	no	alta	insolación
16	no	muy alta	gripe
17	no	muy alta	gripe
18	no	normal	-
19	no	muy alta	gripe
20	si	alta	gripe
21	si	alta	desconocido

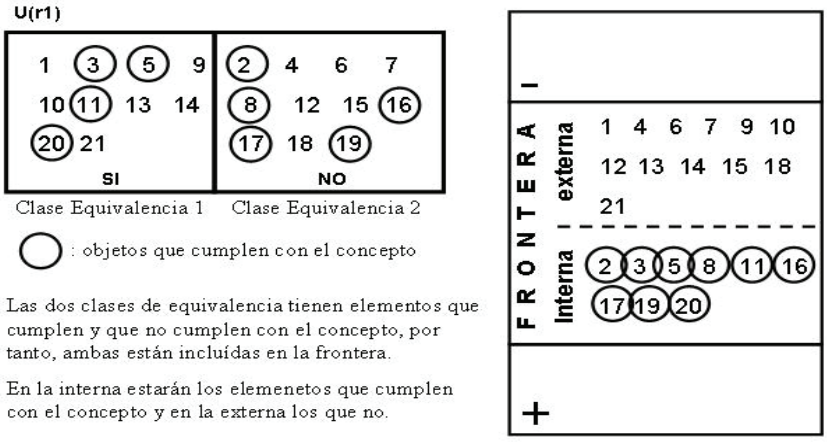

Fig. 2. Partición que estable r_1 sobre U y Frontera de X respecto a r_1

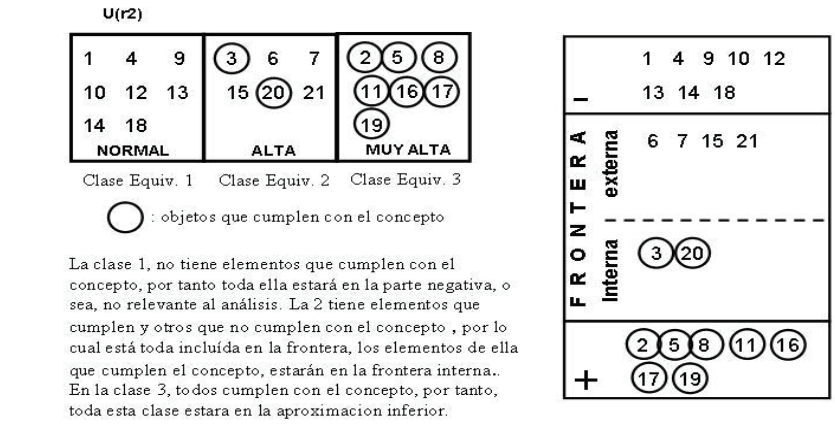

Fig. 3. Partición que estable r_2 sobre U y Frontera de X respecto a r_2

Aplicando el algoritmo, se obtienen las siguientes fronteras internas:

$B_1 = \{2, 3, 5, 8, 11, 16, 17, 19, 20\}$

$B_2 = \{3, 20\}$

El conjunto E calculado que contiene a todos los elementos de U que pertenecen a algún conjunto excepcional no redundante sería el siguiente:

$E = \{3, 20\}$

El grado de excepcionalidad para cada uno de los elementos de E sería:

Grado_de_excepcionalidad (3) = 1

Grado_de_excepcionalidad (20) = 1

Teniendo en cuenta que el valor del grado de excepcionalidad es un valor entre 0 y 1, se puede afirmar que ambos elementos serían considerados *outliers* para cualquier *umbral* μ dado.

Una interpretación de este hecho es que ambos elementos son contradictorios con el elemento 21, que también tiene los mismos síntomas que ellos y sin embargo no se le diagnosticó la GRIPE.

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

Las pruebas realizadas tenían como objetivo fundamental validar la complejidad temporal del algoritmo sobre la base del análisis teórico que se hizo de este parámetro y medir 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 [31]. El *UCI Machine Learning Repository* ofrece una colección de bases de datos (*data sets*) que son usados por la comunidad científica que investiga en temas referidos a *Machine Learning* y *Data Mining* para el análisis empírico de los algoritmos relacionados con estos temas. En particular, las pruebas fueron hechas con un *data set* tomado de este sitio y que contiene datos extraídos del *Census Bureau*

Database de los EE.UU [32] donde hay 48.842 instancias con 14 atributos en los que se mezclan datos continuos y categóricos. En el sitio de la *UCI Machine Learning Repository* aparecen referencias explícitas a más de 50 artículos donde se cita el uso de este *data set.* En el referido sitio pueden apreciarse las características más sobresalientes del mismo y una explicación detallada de sus atributos.

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

8.1 Tiempo de Ejecución

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

Los resultados que se muestran en la Fig. 4. permiten concluir que la dimensionalidad del *data set* (cantidad de columnas de la tabla) no influye en el tiempo de ejecución. Puede verse además que un aumento de la misma no representa un problema para que el método se ejecute correctamente. Los resultados en cuanto a los tiempos de ejecución alcanzados corroboran lo expresado en el análisis teórico realizado para el cálculo de la complejidad temporal del algoritmo.

Los resultados mostrados en la gráfica de la Fig. 5. reflejan los tiempos de ejecución alcanzados por el algoritmo haciendo variar la cantidad de filas del *data set*. Se puede apreciar que la variación considerada oscila entre 5.000 y 30.000 filas, obteniéndose en todos los casos tiempos de ejecución razonables para el volumen de información procesado.

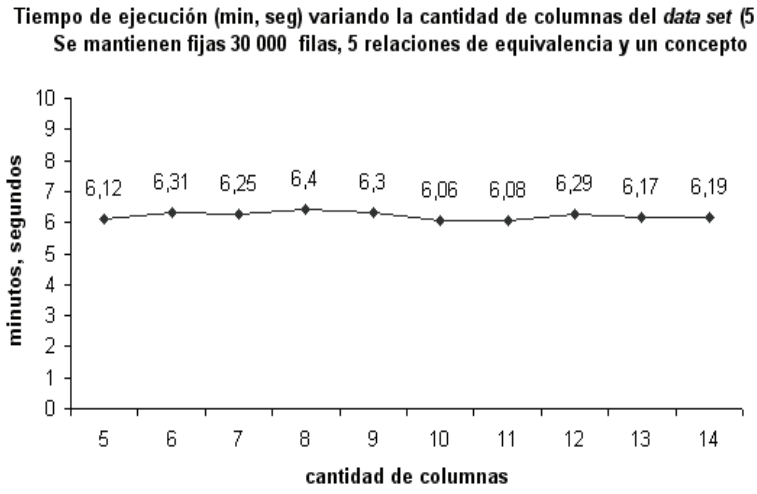

Fig. 4. Variando número de columnas del *data set*

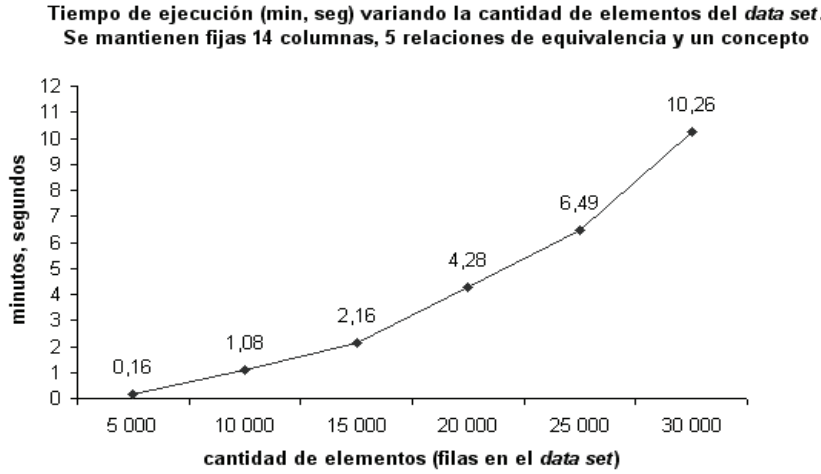

Fig. 5. Variando la cardinalidad del *data set*

En la Fig. 6. se muestra la dependencia del tiempo de ejecución con respecto a la cantidad de relaciones de equivalencia. Teóricamente fue demostrado que dicha dependencia es cuadrática y en esta gráfica podemos ver que para valores pequeños de m, dicha dependencia se comporta casi lineal. Es decir, todo indica que las constantes definen una parábola muy abierta, por tanto, para valores no muy grandes de m, que es lo más usual, se garantiza la casi linealidad de la complejidad temporal del algoritmo con respecto a dicho parámetro.

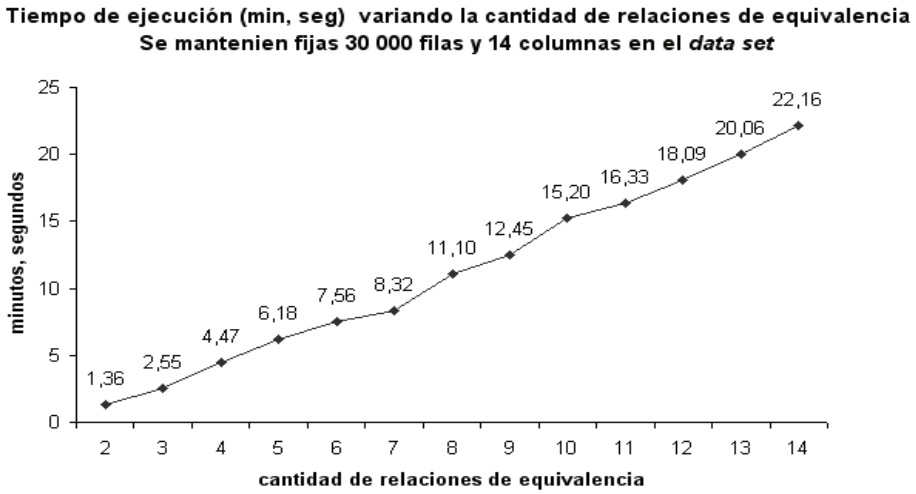

Fig. 6. Variando el número de relaciones de equivalencia

8.2 Detección

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

- Los individuos de la tabla que fueron objeto de estudio fueron 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']
```

Nótese que cualquier elemento que cumpla el concepto y pertenezca a clase *x*.1 (*x* = 1, 2, 3, 4) es contradictorio por la relación *rx*, teniendo en cuenta que los individuos sujetos al análisis son niños entre 1 y 10 años.

El conjunto de *outliers* con los cuales fue *bombardeado* el *data set* se muestra en la Tabla 2.

Los valores de los restantes atributos para estos elementos son irrelevantes al análisis, debido a que las relaciones de equivalencia con las que se trabaja solo toman en consideración los atributos *workclass*, *education*, *marital status* y *occupation*. Los valores de los atributos resaltados en **negrita** e *itálica* son contradictorios para niños con edades entre 1-10. Nótese que en el conjunto de *outliers* introducidos, 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 atributos y éstos son precisamente los elementos más contradictorios.

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***

La Fig. 7. muestra la cantidad de outliers detectados para diferentes valores del umbral de detección μ en diferentes pruebas que fueron realizadas.

La diferencia entre las pruebas que se ilustran en Fig. 7 queda establecida por la cantidad de *outliers* introducidos en el data set. Como aspecto a destacar en los resultados alcanzados cabe señalar que siempre en el conjunto de *outliers* detectados se encontraron los *outliers* introducidos. Esto se cumple tanto cuando la cantidad detectada fue mayor que la cantidad introducida así como cuando el número de outliers detectados fue menor que dicha cantidad.

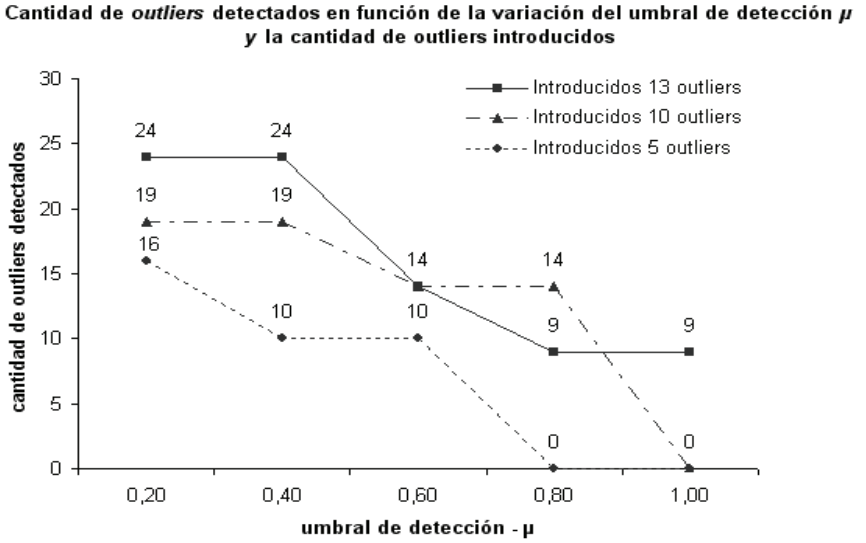

Fig. 7. Resultados de las pruebas de detección

Lo anterior refleja el nivel de eficiencia que puede alcanzar el algoritmo en cuánto a detección. Por su parte la variación del valor del *umbral* μ conlleva un cierto refinamiento en la detección, aunque en algunos casos no se logra. Esto se manifiesta en algunos estancamientos que se aprecian en las gráficas. En otras ocasiones se deja bruscamente de detectar *outliers*, lo cual se evidencia en los pronunciados descensos a cero que ocurren. La causa de este escaso refinamiento parece estar provocada por el carácter determinista del método en lo que respecta a clasificación que, por estar basado en el modelo básico de *Rough Sets*, hereda sus limitaciones. Este hecho hace suponer que, al permitirse un cierto grado de desclasificación, se podrá lograr en algunos casos una mayor calidad en la detección, obteniendo finalmente como *outliers* a los elementos más contradictorios.

9 Conclusiones

En este trabajo se ha propuesto un método de detección de *outliers* cuyas principales aportaciones son las siguientes: el método es simple en cuanto a su planteamiento teórico. La definición de *outliers* propuesta es intuitiva, simple y computacionalmente factible para grandes *data set*s. El algoritmo propuesto es eficiente para hacer minería de *outliers* y tiene como base conceptual la teoría de RS. Su enfoque es original y novedoso pues no existen, al menos en la bibliografía revisada, antecedentes con un planteamiento similar. No cae dentro de ninguna de las categorías que la bibliografía recoge para clasificar los métodos de detección según el marco teórico en el cual se basan los mismos.

En la actualidad, los conjuntos de datos del mundo real y sus entornos presentan un amplio rango de dificultades y esto limita la efectividad del uso de determinados métodos de detección. Entre algunos de los problemas que pueden señalarse en este sentido se encuentra el hecho de que los conjuntos de datos (*data sets*) pueden ser dinámicos, imponiendo la necesidad de usar algoritmos eficientes en cuanto a su complejidad temporal. Por ejemplo, la mayoría de los métodos basados en distancias son de orden al menos cuadrático con respecto al número de elementos en el *data set*, lo cual quizás sea inaceptable si el *data set* es muy grande o dinámico.

El algoritmo que se propone es lineal con respecto a la cardinalidad del universo de datos sobre el cual se aplica y es cuadrático respecto al número de relaciones de equivalencia usadas para describir dicho universo pero dicho número representa una constante y su valor suele ser significativamente menor que la cardinalidad del universo.

Resulta importante destacar que el método es aplicable a datos en forma tabular. La tabla es la estructura de datos del Modelo Relacional. La misma debe estar, como mínimo, en 1ª forma normal para garantizar que no haya redundancias y sus atributos deben ser monovaluados, de lo contrario, entrarían en contradicción con la esencia del método pues no existiría la posibilidad de establecer a partir de ellos relaciones de equivalencia. Lo anteriormente expresado decanta el ámbito de aplicación del problema.

Otro elemento que puede limitar la efectividad de un método es la naturaleza de los datos. A diferencia de un gran número de métodos que presentan dificultades para su

aplicación atendiendo a este aspecto, el método propuesto es aplicable tanto a datos continuos como a datos discretos. El hecho de que los *data sets* puedan contener una mezcla de tipos de atributos (p.ej., atributos continuos y categóricos mezclados) no es una limitación para la aplicación del algoritmo propuesto. Sin embargo, en otros métodos, especialmente los basados en distancias y en densidades, sí lo es.

Los métodos estadísticos se centran básicamente en la detección de *outliers* sobre datos univariados. En ellos se requiere un conocimiento a priori de la distribución de los datos. En estos casos el usuario tiene que modelar los datos usando una distribución estadística y los *outliers* son determinados en función de cómo aparecen en relación al modelo postulado. El principal problema de esta aproximación estriba en el número de situaciones que pueden existir y el usuario quizás no tenga el conocimiento suficiente de la distribución de los datos. La aproximación que proponemos no requiere ningún conocimiento a priori de la distribución de los datos.

La principal limitación del método propuesto radica en que la concepción del método se basa en el modelo básico de *Rough Sets* al cual se le critica su incapacidad para modelar información incierta. La clasificación con un grado controlado de incertidumbre, o un posible error de clasificación está fuera del alcance de este modelo. Sin embargo, en la práctica, poder admitir algún nivel de incertidumbre en el proceso de clasificación puede conllevar a una comprensión más profunda y a una mejor utilización de las propiedades de los datos analizados. La definición estándar de inclusión de conjunto tenida en cuenta en el modelo actual es demasiado rigurosa para modelar una inclusión de conjuntos *casi* completa. Teniendo en cuenta esta limitación, en la actualidad estamos trabajando en una nueva versión del algoritmo basada en el Modelo de *Rough Sets* de Precisión Variable propuesto por el profesor Ziarko en 1993 [33] como solución a las dificultadas fundamentales que se le señalan al modelo básico de *Rough Sets* el cual se considera como caso particular del modelo de precisión variable.

Referencias

1. Hawkins, D. (1980). Identification of outliers. Chapman and Hall, Reading.
2. Barnett, V., & Lewis, T. (1994). Outliers in Statistical Data, 3rd edn. John Wiley & Sons.
3. Graham, W., Baxter, R. A., He, H. X., Hawkins, S., & Gu, L. (2002). A comparative study of RNN for Outlier Detection in Data Mining. IEEE International Conference on Data-Mining (ICDM'02). CSIRO Technical Report CMIS-02/102. Maebashi City, Japan.
4. Angiulli, F., Greco, G., & Palopoli, L. (2006). Outlier Detection by Logic Programming. ACM Transaction on Computational Logic, Vol. V , 1-50.
5. Stomoimenova, E., Mateev, P., & Dobreva, M. (2006). Outlier detection as a method for Knowledge extraction from digital resources. Institute of Mathematics and Informatics, Bulgarian Academy of Sciences.
6. Li, S., Lee, R., & Lang, S.-D. (2006). Detecting outliers in interval data. ACM Southeast Regional Conference, (pages: 290-295).
7. He, Z., Xu, X., Huang, J. Z., & Deng, S. (2004). Mining class outlier: concepts, algorithms and applications in CRM. Expert System with Applications.

8. Lin, S., & Brown, D. E. (2001). Outlier-based Data Association:Combining OLAP and Data Mining. Technical Report. Dept. of Systems Engineering Univ. of Virginia.
9. Knorr, E., Ng, R., & Tucakov, T. (2000). Distance-based outliers: Algorithms and Applications. VLDB Journal, 8 (3 and 4) , 237-253.
10. Cramer J, A., Shah S, S., Battaglia T, M., Banerji S, N., Obando L, A., & Booksh K, S. (2004). Outlier detection in chemical data by fractal analysis. Journal of Chemometrics; Volume 18, Issue 7-8 , 317-326.
11. 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, 2005.
12. Hodge, V. J., & Austin, J. (2004). A Survey of Outlier Detection Methodologies. Artificial Intelligence Review 22 , 85-126.
13. Knorr, E., & Ng, R. (1998). Algorithms for mining distance-based outliers in large datasets. In Proceedings of theVLDB98, (pages: 392-403).
14. Knorr, E., & Ng, R. (1999). Finding intentional knowledge of distance-based outliers. In Proceedings of the VLDB99, (pages: 211-222).
15. Schwabacher, M., & Bay, S. D. (2003). Mining distance-based outliers in near lineal time with randomization and a simple pruning rule. In Proc. of 9th annual ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
16. Tang, J., Chen, Z., Fu, A., & Cheung, D. (2002). A Robust Outlier Detection Scheme in Large Data Sets. 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Taipei, Taiwan.
17. Yamanishi, K., & Takeuchi, J. (2001). Discovering outlier filtering rules from unlabeled data- combining a superviser learner with am unsuperviser learner. In Proceedings of the KDD01 , (pages: 389-394).
18. Theodore, J., Ivy, K., & Raymong, T. N. (1998). Fast Computation of 2d depth contours. In ACM SIG KDD, (pages: 224-228).
19. Schwabacher, M., & Bay, S. D. (2003). Mining distance-based outliers in near lineal time with randomization and a simple pruning rule. In Proc. of 9th annual ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
20. Ren, D., Wang, B., & Perrizo, W. (2004). RDF: A density-based Outlier Detection Method using Vertical Data Representation. In Proceedingas of the Fourth IEEE Internatinal Conference on Data Mining (ICDM'04), Computer Science Department. North Dakota State University, Fargo, ND 58105, USA.
21. He, Z., Deng, S., & Xu, X. (2002). Outlier detection integrating semantic knowledge. WAIM'02.
22. Petrovsky, M. (2003). A Hybrid Method for Patterns Mining and Outliers Detection in the Web Usage Log. AWIC'03, (pages: 318-328).
23. Kandel, A., & Last, M. (2001). Automated detection of outliers in real-world data. In Proceedings of the Second International Conference on Intelligent Technologies. Bangkok, Thailand.
24. Sun, P., & Chawla, S. (2006). Outlier Detection: Principles, Techniques and Application. PaKdd 2006. Singapore.
25. Pawlak, Z. (1991). Rough Sets: Theoretical Aspects of Reasoning About Data. s.l.: Spinger.
26. 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.
27. 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 C. Society.

28. 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.
29. Rokach, L. (2008). An evolutionary algorithm for constructing a decision forest: Combining the classification of disjoints decision trees. Wiley InterScience. International Journal of Intelligent Systems , Volume 23, Issue 4, Pages: 455-482.
30. 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.
31. University of California Irvine. Center for Machine Learning and Intelligent Systems - http://cml.ics.uci.edu – Ultimo acceso: 30/06/2009
32. USA Census Bureau Database - http://www.census.gov/ftp/pub/DES/www/welcome.html – Ultimo acceso: 30/06/ 2009
33. Ziarko, W. (1993). Variable Precision Rough Set Model. s.l. : J. Comput. Syst. Sci.