Revista Cubana de Ciencias Informáticas
Vol. 11, No. 2, Abril-Junio, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 61 -72

Tipo de artículo: Artículo original
Temática: Tecnologías de bases de datos
Recibido: 17/05/2016 | Aceptado: 09/02/2017

Algoritmo OneR. Su aplicación en ensayos clínicos

OneR Algorithm. Its application on Clinical Trials

Yanet Cardoso García [1*], Luis Arza Valdés [2]

[1] Centro de Tecnologías de Gestión de Datos. Universidad de las Ciencias Informáticas. Carretera a San Antonio de los Baños, Km 2½, Reparto Torrens, La Habana, Cuba. CP.: 19370. ycardosog@uci.cu
[2] Empresa Comercializadora y de Servicios a la Salud. Avenida 43 % 44 y 48, La Ceiba. Playa. La Habana, Cuba. luis8206@gmail.com

[*] Autor para correspondencia: ycardosog@uci.cu

Resumen

Los Ensayos Clínicos constituyen una etapa determinante en el desarrollo de cualquier producto médico. El gran cúmulo de datos que generan y la información oculta en estos continúan siendo una importante disyuntiva para la comunidad científica. La minería de datos se convierte en una de las soluciones para el problema en cuestión. En el presente trabajo se aplica un algoritmo de reglas de inducción, el OneR, con el objetivo de mostrar patrones desde los datos contenidos en el mercado de datos del producto LeukoCIM, en fase de Ensayo Clínico por parte de los especialistas del Centro de Inmunología Molecular. La aplicación del algoritmo se llevó a cabo a través del proceso de minería de datos, guiado por la metodología CRISP-DM e integrado en el Sistema Gestor de Bases de Datos PostgreSQL. Finalmente se obtuvo un conjunto de reglas de clasificación asociadas al grado del evento adverso que presenta la enfermedad base de los pacientes, evidenciando la viabilidad de la utilización del algoritmo OneR en esta sensible rama de la sociedad.

Palabras clave: algoritmo OneR, Ensayos Clínicos, minería de datos

Abstract

Clinical Trials constitute a critical stage in the development of any medical product. The large body of data generated and the information hidden in them, remain as one of the dilemmas for the scientific community. Data mining becomes one of the solutions to the problem at hand. In this paper, a OneR algorithm for induction rules is applied, with the aim of showing patterns from the data contained in the LeukoCIM product data market, in clinical trial phase by the specialists of the Molecular Immunology Center. The application of the algorithm was carried out

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 2, Abril-Junio, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 61 -72

through the data mining process, guided by the Cross-Industry Standard Process for Data Mining methodology and integrated into the PostgreSQL Database Management System. As a result, it was obtained a set of classification rules related to the level of the adverse event presented in the base disease of patients, demonstrating the feasibility of using the OneR algorithm in this sensitive sector of the society.

Keywords: *Clinical Trials, data mining, OneR algorithm*

Introducción

En la sociedad actual, el gran volumen de datos e información almacenada es un elemento determinante en el desarrollo del conocimiento y, por ende, en la construcción del mundo moderno. La máxima "la información es poder" ha marcado un cambio profundo en la forma en que las personas interactúan, incrementando el debate, los conflictos y los puntos de vista hacia la evolución de las Tecnologías de la Información y las Comunicaciones (TIC).

Para tener una idea más acertada de la anterior afirmación, en el año 2000 se generaron 800.000 petabytes(PB) de datos almacenados y se espera que esta cifra alcance los 35 zettabytes(ZB) en el año 2020. Algunas empresas generan terabytes de datos cada hora de cada día del año, estando inundadas de ellos (IBM, 2012). La necesidad de analizar este volumen de datos y explotarlos de forma eficaz, constituye un objetivo de cualquier organización; tal es el caso de las instituciones científicas que son beneficiadas con el uso de las tecnologías, siendo el Centro de Inmunología Molecular (CIM) de Cuba una de ellas.

La minería de datos es una herramienta fundamental para el análisis y descubrimiento de conocimiento a partir de datos, permitiendo aglutinar una variedad de metodologías analíticas, proporcionando un marco conceptual y metodológico para el abordaje del análisis de señales en distintas disciplinas. Sin embargo, cada campo de aplicación presenta desafíos propios que deben ser abordados particularmente desde la racionalización de los conceptos específicos del ámbito, en este caso el biotecnológico.

El CIM se destaca por la utilización y desarrollo de la biología molecular con el fin de estudiar diversos tipos de cáncer, así como la elaboración de vacunas contra dicha enfermedad y anticuerpos monoclonales[1]. Para la aprobación, prueba e introducción de medicamentos en el mercado, se realizan Ensayos Clínicos que consisten en *"... cualquier investigación en seres humanos dirigida a descubrir o verificar los efectos clínicos, farmacológicos u otros efectos*

[1] Sustancia producida en el organismo animal por la presencia de un antígeno, contra cuya acción reacciona específicamente.

farmacodinámicos de un producto en investigación..." (Ministerio de Salud Pública, 2009). Estos cuentan de forma general con cuatro fases, donde:

– La fase I incluye los primeros estudios que se realizan en seres humanos denominándose fase de estudios de farmacología humana.

– La fase II tiene como objetivo proporcionar información preliminar sobre la eficacia del producto y establecer la relación dosis-respuesta; son estudios terapéuticos exploratorios.

– Los ensayos clínicos de fase III evalúan la eficacia y seguridad del tratamiento experimental en las condiciones de uso habituales y con respecto a las alternativas terapéuticas disponibles para la indicación estudiada. Se trata de estudios terapéuticos de confirmación.

– La fase IV se realiza después de la comercialización del fármaco para estudiar condiciones de uso distintas de las autorizadas, como nuevas indicaciones, y la efectividad y seguridad en la utilización clínica diaria (Bakke, Carné, García, 1994).

Como se aprecia, cada fase tiene objetivos bien definidos para ir valorando la eficacia del producto. Esto en gran medida transita por pruebas que van desde pocos sujetos hasta cientos de miles, generando un cúmulo de datos inmenso relacionado con las características y evolución ante la administración del fármaco de cada individuo; datos que deben ser analizados desde lo específico hasta lo general.

Uno de los productos de investigación del CIM es el fármaco LeukoCIM, el cual transitó por cuatro Ensayos Clínicos, siendo sus datos almacenados en un mercado de datos, con el objetivo de extraer conocimiento relacionado con el medicamento en cuestión.

De acuerdo a las características de los datos y los disímiles algoritmos de minería que se utilizan con este fin, y a partir de las necesidades de la institución científica, se decide aplicar un algoritmo descriptivo que permitiera a través de reglas de inducción representar hipótesis más comprensibles para los especialistas de esta entidad y extraer patrones o modelos de los datos. Se consideró entonces, aplicar el OneR, el cuál es uno de los algoritmos clasificadores más sencillos y rápidos; dado que simplemente identifica el atributo que mejor explica la clase de salida (Witten, 2005).

Precisamente para lograr entender el gran cúmulo de datos que se generó de estos Ensayos Clínicos y contribuir con la toma de decisiones de los investigadores del CIM, se determinó como objetivo fundamental de este trabajo aplicar el algoritmo OneR como parte del complejo proceso de extracción de conocimiento oculto en los datos.

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 2, Abril-Junio, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 61 -72

Materiales y métodos

En la presente investigación se profundizó sobre el emergente y dinámico campo de investigación denominado: Descubrimiento de conocimiento en Bases de Datos (KDD, en inglés *Knowledge Discovery in Data Bases*). Este presenta una secuencia de pasos o fases que son: preparación de los datos (selección y transformación), minería de datos, evaluación, interpretación y toma de decisiones.

Una de las fases más importantes dentro de este proceso es la minería de datos que integra técnicas de análisis de datos y extracción de modelos (Usama, 1996), además se define como el análisis de archivos que trabaja a nivel de conocimiento con el fin de descubrir patrones, relaciones, reglas, asociaciones o incluso excepciones útiles para la toma de decisiones (Rodríguez, 2009).

Técnicas de minería de datos

Las técnicas de minería de datos constituyen un enfoque conceptual y habitualmente son implementadas por varios algoritmos (Molina; García, 2006). Estas pueden clasificarse en dependencia de su utilidad en técnicas de predicción, asociación, agrupamiento (clustering) y clasificación.

- ✓ **Las técnicas de predicción** permiten obtener pronósticos de comportamientos futuros a partir de los datos recopilados.
- ✓ **Las técnicas de reglas de asociación** permiten establecer las posibles relaciones o correlaciones entre distintas acciones o sucesos aparentemente independientes, pudiendo reconocer como la ocurrencia de un suceso o acción puede inducir o generar la aparición de otros.
- ✓ **Las técnicas de agrupamiento** concentran datos dentro de un número de clases preestablecidas o no, partiendo de criterios de distancia o similitud, de manera que las clases sean similares entre sí y distintas con las otras clases.
- ✓ **Las técnicas de clasificación** consisten en definir una serie de clases donde se puedan agrupar los diferentes casos. Dentro de este grupo se encuentran los árboles de decisión y las reglas de inducción (Castillo, 2014).

Técnicas de clasificación

Dentro de las técnicas de clasificación se encuentra el árbol de decisión y las reglas de inducción, siendo esta última la utilizada durante la investigación, la cual permite la generación de reglas a partir de los datos de entrada. La información de entrada será un conjunto de casos donde se ha asociado una clasificación o evaluación a un conjunto de variables o atributos (Ruiz, 2008).

Las reglas permiten expresar disyunciones de manera más fácil que los árboles y tienden a preferirse con respecto a estos por mostrar "partes" de conocimientos relativamente independientes. Representan funciones que establecen una relación entre los ejemplos (descritos mediante un conjunto de rasgos) y las clases de decisión. Se expresan de la forma If P then Q, donde P es la parte condicional formada usualmente por una conjunción de condiciones elementales, y Q es la parte de decisión que asigna un valor de decisión (clase) a un objeto que cumpla la condición. Las reglas constituyen patrones que establecen una dependencia entre los valores de los atributos de condición en P y el valor de decisión Q (Filiberto; Bello; Caballero, 2011).

Algunos de los beneficios de las reglas de inducción es que son las representaciones de hipótesis más perceptibles para el hombre y el formalismo más notorio de representación del conocimiento, de ahí que sean muy utilizadas en aplicaciones médicas. Es por ello que en la investigación se decide aplicar esta técnica, debido a que es muy descriptiva.

Algoritmos

Existen un grupo de algoritmos que utilizan como base los datos de tipo nominal, entre ellos se encuentran: tabla de decisión, ID3, C4.5, PART y el OneR; siendo este último uno de los algoritmos implementados para PostgreSQL dentro de las reglas de inducción. El OneR está basado en el algoritmo ID3, donde la meta principal consiste en adquirir las reglas de clasificación directamente desde el conjunto de datos de entrenamiento (Gasparovica; Aleksejeva, 2010). Es por tanto el algoritmo de clasificación seleccionado ya que es simple y efectivo, de uso frecuente en el aprendizaje de máquinas, utiliza un único atributo para la clasificación, el cual es el de menor porciento de error y se obtiene un conjunto de reglas.

Si existen atributos numéricos, busca los umbrales para hacer reglas con mejor tasa de aciertos (Witten, 2005). Al utilizar un clasificador, su precisión y fiabilidad depende principalmente de los casos clasificados correctamente a partir del número total de elementos (fase de evaluación e interpretación de resultados).

Además permite analizar atributos con valores nominales, pues en el negocio donde se va a aplicar, los valores de los atributos seleccionados no presentan valores continuos, ni se trabaja con fechas, horas y las tuplas con valores desconocidos se eliminaron previamente.

Herramienta seleccionada para aplicar el algoritmo OneR.

Para aplicar el algoritmo OneR existen diversas herramientas; algunas son independientes del Sistema Gestor de Bases de Datos (SGBD) y otras son nativas de un SGBD específico. Dentro de las herramientas nativas del gestor se

pueden encontrar Oracle Server Data Mining y Oracle Data Mining. Además están las independientes del gestor como son Clementine de SPSS, Weka (*Waikato Environment for Knowledge Analysis*) y SAS Enterprise Miner (Rodríguez, 2009).

Teniendo en cuenta que es muy beneficioso aplicar el algoritmo en el mismo SGBD, se utilizó el PostgreSQL con el auxilio de una extensión que contiene la implementación de varios algoritmos de minería de datos. De esta forma el proceso es menos engorroso ya que el tiempo para la preparación y vinculación de los datos es mínimo, acortando el tiempo de respuesta de los análisis (Robles, 2012).

También se hace uso de la herramienta pgAdmin III ya que es un cliente gráfico para el SGBD PostgreSQL, es multiplataforma y permite gestionar todos los objetos de la base de datos e incluye un editor SQL con resaltado de sintaxis (PostgreSQL, 2011).

Metodología aplicada

Para realizar la extracción de conocimiento útil a partir de los datos almacenados existen metodologías que permiten llevar a cabo el proceso de minería de datos en forma sistemática y no tribial. Dentro de las existentes se destacan SEMMA y CRISP-DM. La primera se centra en las características técnicas del desarrollo del proceso y la segunda realiza un análisis global del negocio al cual se le va a aplicar, por lo que la investigación se desarrolla bajo esta última. CRISP-DM (de las siglas en inglés Cross-Industry Standard Process for Data Mining) contiene 6 fases encargadas de analizar el problema, analizar los datos, prepararlos, realizar la modelación, la evaluación y la explotación.

En la primera fase, comprensión del problema, esta metodología propone el estudio de los objetivos y requerimientos del proyecto para definir un plan preliminar para alcanzar las metas. En la fase de comprensión de los datos se recolectan los que se van a utilizar, se describen y se verifica la calidad de los mismos. La fase de preparación incluye la integración, selección, limpieza y transformación de los datos. En la fase de modelación se seleccionan las técnicas y algoritmos a utilizar en dependencia del objetivo a resolver, luego se procede a evaluar comprobando que se cumplan los objetivos del negocio y por último se le presentan a los especialistas informes sobre el conocimiento extraído.

Implementación e integración del algoritmo OneR

Para la implementación del algoritmo OneR se asumió el resultado de la Tesis de Maestría de Yadira Robles en el año 2012, donde utilizó el lenguaje Pl/pgsql en la realización de una función que toma como entrada el nombre de la tabla

y la clase sobre la cual se va a realizar el análisis, y devuelve como resultados un conjunto de reglas para los atributos con la menor cantidad de errores.

La integración de dicho algoritmo se realizó mediante una extensión creada para PostgreSQL. La extensión contiene dos archivos, uno que define las características de la misma y el otro los objetos SQL que fueron agregados (Robles, 2012). Ambos archivos fueron ubicados dentro del directorio de la instalación "/usr/share/postgresql/9.3/extension/".

Resultados y discusión

Una vez analizado el algoritmo seleccionado perteneciente a las técnicas de clasificación, las herramientas que lo soportan, la integración a estas, y la metodología a utilizar se exponen los resultados alcanzados durante la investigación.

Comprensión del negocio

Los objetivos del negocio están enmarcados a:

- ✓ Conocer las relaciones que se establecen entre las características de los pacientes.
- ✓ Determinar cuáles son las características de las personas que pueden influir en el grado los eventos adversos.

Como objetivo de la minería de datos se identificó:

- ✓ Obtener reglas que determinen el parámetro más significativo, así como la influencia que tiene en el grado de los eventos adversos, analizando la edad, el sexo, el peso, la raza y la enfermedad base del paciente.

Comprensión de los datos

La comprensión de los datos está relacionada con la recolección y descripción de la información que contiene la base de datos resultado de la tesis de diploma realizada en el 2013 por la autora del presente trabajo. La información recopilada fue obtenida de una única fuente, que de manera centralizada es la que contiene todos los datos de los pacientes que participaron en los Ensayos Clínicos del producto LeukoCIM. Esta base de datos cuenta con 33 tablas contenidas en tres esquemas (Cardoso, 2013).

A continuación, se describen los atributos significativos para darle solución a los objetivos propuestos en el epígrafe Comprensión del negocio.

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 2, Abril-Junio, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 61 -72

Tabla 1. Resumen descriptivo de la información relevante recopilada

Nombre del atributo	Tipo de dato	Descripción	Nombre de la tabla
dk_codigo_paciente	character varying	Almacena el identificador de cada paciente	hech_ninno_adulto_sida_neutrop_ea
raza_descripcion	character varying	Almacena la raza del paciente	dim_raza
sexo_descripcion	character varying	Almacena el sexo del paciente	dim_sexo
peso_descripcion	character varying	Almacena el peso del paciente	dim_peso
grado_evento_adverso	character varying	Almacena el grado del evento adverso del paciente	dim_grado_evento_adverso
enfermedad_base	character varying	Almacena la enfermedad inicial del paciente	dim_enfermedad_base

Explorar y verificar la calidad de los datos

Teniendo en cuenta que los datos utilizados para la aplicación del algoritmo fueron extraídos de un mercado de datos, estos se encuentran estructurados, integrados y sin ruidos. Se comprobó que de las 145 688 tuplas contenidas en 33 tablas no existían valores nulos ni duplicados.

Preparación de los datos

Selección de los datos

De los atributos vistos en el epígrafe Comprensión de los datos, se excluyó de la selección el denominado dk_codigo_paciente, que se empleó solo como enlace entre las tablas. Este presenta gran variabilidad ya que cada instancia es única.

Limpieza y transformación de los datos

Para la limpieza de los datos se realizó un perfilado a los mismos, lo que permite comprender su contenido, estructura, calidad y dependencias.

Para un mayor entendimiento en el trabajo con los valores de los atributos se realizaron las siguientes transformaciones:

- ✓ Los valores del sexo que aparecían con 1 y 2 se sustituyeron por Masculino y Femenino respectivamente.
- ✓ Los valores de la raza que aparecían con 1, 2, 3 y 4 se sustituyeron por Blanca, Negra, Mestiza y Amarilla respectivamente.
- ✓ Los valores del grado del evento adverso según la Organización Mundial de la Salud (OMS) que aparecían

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 2, Abril-Junio, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 61 -72

con 1, 2, 3, 4, 5, 6 y 7 se sustituyeron por Normal, Ligero, Moderado, Severo, Que amenaza la vida, Que causa la muerte y Muy severo respectivamente.

Integración de datos

Los campos seleccionados se encontraban en diferentes tablas por lo cual se realizó una consulta uniendo los atributos necesarios para dar cumplimiento al objetivo del negocio, creándose una nueva tabla como resultado de una vista que contiene 476 registros con un total de cinco campos.

Modelado

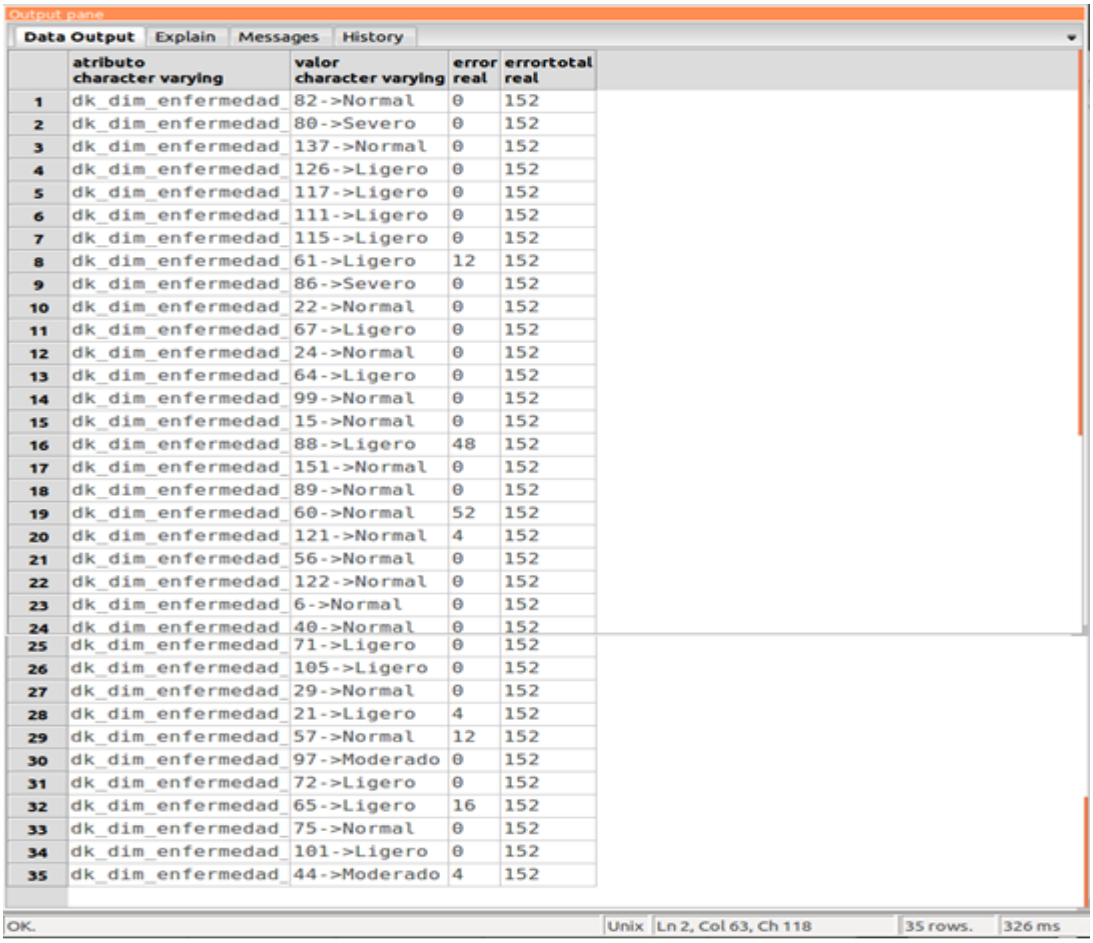
La selección de las técnicas de minería de datos y el algoritmo a emplear depende de los objetivos de minería de datos propuestos en la fase de comprensión del negocio. La técnica seleccionada para darle cumplimiento a este objetivo es la de reglas de inducción y de ellas el algoritmo OneR.

Para obtener las reglas mediante dicho algoritmo integrado al SGBD PostgreSQL, se realizó la siguiente consulta “SELECT * FROM "1r" (' pgday ',' grado_evento_adverso ') as (atributo varchar, valor varchar, error real, errortotal real)”, haciendo referencia a la función implementada pasándole como parámetros el nombre de la tabla y el atributo clase, el cual va a ser por el que se va a analizar la tabla. (Ver Figura 1).

Al aplicar el algoritmo OneR se obtuvieron estos resultados los cuales se interpretan de la siguiente forma:

1. Si la enfermedad base del paciente es Linfoma de Hodgkin la influencia del grado del evento adverso es Normal.
2. Si la enfermedad base del paciente es Linfoma Cutáneo de Células T la influencia del grado del evento adverso es Severo.
3. Si la enfermedad base del paciente es Síndrome Dismielopoyético la influencia del grado del evento adverso es Normal.
4. Si la enfermedad base del paciente es Osteosarcoma de Fémur la influencia del grado del evento adverso es Ligero.
5. Si la enfermedad base del paciente es Neoplasia de Mama con Metástasis Pulmonar la influencia del grado del evento adverso es Ligero.

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 2, Abril-Junio, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 61 -72



Output pane

Data Output | Explain | Messages | History

	atributo character varying	valor character varying	error real	errortotal real
1	dk_dim_enfermedad_	82->Normal	0	152
2	dk_dim_enfermedad_	80->Severo	0	152
3	dk_dim_enfermedad_	137->Normal	0	152
4	dk_dim_enfermedad_	126->Ligero	0	152
5	dk_dim_enfermedad_	117->Ligero	0	152
6	dk_dim_enfermedad_	111->Ligero	0	152
7	dk_dim_enfermedad_	115->Ligero	0	152
8	dk_dim_enfermedad_	61->Ligero	12	152
9	dk_dim_enfermedad_	86->Severo	0	152
10	dk_dim_enfermedad_	22->Normal	0	152
11	dk_dim_enfermedad_	67->Ligero	0	152
12	dk_dim_enfermedad_	24->Normal	0	152
13	dk_dim_enfermedad_	64->Ligero	0	152
14	dk_dim_enfermedad_	99->Normal	0	152
15	dk_dim_enfermedad_	15->Normal	0	152
16	dk_dim_enfermedad_	88->Ligero	48	152
17	dk_dim_enfermedad_	151->Normal	0	152
18	dk_dim_enfermedad_	89->Normal	0	152
19	dk_dim_enfermedad_	60->Normal	52	152
20	dk_dim_enfermedad_	121->Normal	4	152
21	dk_dim_enfermedad_	56->Normal	0	152
22	dk_dim_enfermedad_	122->Normal	0	152
23	dk_dim_enfermedad_	6->Normal	0	152
24	dk_dim_enfermedad_	40->Normal	0	152
25	dk_dim_enfermedad_	71->Ligero	0	152
26	dk_dim_enfermedad_	105->Ligero	0	152
27	dk_dim_enfermedad_	29->Normal	0	152
28	dk_dim_enfermedad_	21->Ligero	4	152
29	dk_dim_enfermedad_	57->Normal	12	152
30	dk_dim_enfermedad_	97->Moderado	0	152
31	dk_dim_enfermedad_	72->Ligero	0	152
32	dk_dim_enfermedad_	65->Ligero	16	152
33	dk_dim_enfermedad_	75->Normal	0	152
34	dk_dim_enfermedad_	101->Ligero	0	152
35	dk_dim_enfermedad_	44->Moderado	4	152

OK. | Unix | Ln 2, Col 63, Ch 118 | 35 rows. | 326 ms

Figura 1: Resultados del análisis realizado mediante el algoritmo OneR integrado a PostgreSQL

El resto de los resultados se interpretan de la misma manera, teniendo en cuenta que el número que antecede el grado del evento adverso se refiere a la llave primaria del atributo dk_dim_enfermedad_base_id.

Evaluación e implementación

Este proceso de evaluación de los resultados, se realizaba por parte de los especialistas del CIM a través de los Cuadernos de Recogida de Datos, que posteriormente eran insertados en un sistema llamado EpiData y finalmente lo exportaban al formato Excel provocando márgenes de errores elevados y tiempo considerable. Es por ello que se desarrolla el mercado de datos con el mismo objetivo del negocio relacionado con el descubrimiento de patrones ocultos en los datos. Con la aplicación del algoritmo OneR se logra clasificar el grado del evento adverso, basado en las relaciones que se establecen entre los atributos seleccionados. Por lo que puede considerarse el modelo como aceptado, desde el punto de vista analítico, para apoyar la toma de decisiones médicas y administrativas del CIM.

Conclusiones

En la presente investigación se aplicó el algoritmo OneR a través de la integración de la extensión realizada para el SGBD PostgreSQL. Se analizó la técnica de clasificación y su vínculo con el algoritmo en cuestión. Con la guía de la metodología CRISP-DM se llevó a cabo el proceso de minería de datos.

La información descubierta a partir de la aplicación del algoritmo OneR, permite a los directivos del Centro de Inmunología Molecular emprender las acciones y determinar, si así lo estiman conveniente, una estrategia a seguir en la administración del producto LeukoCIM que posibilite elevar el nivel de vida de los pacientes que lo consumen.

Referencias

BAKKE OM, CARNÉ X, GARCÍA ALONSO F. Investigación y desarrollo de nuevos fármacos. En: Bakke OM, Carné X, García-Alonso F (ed). Ensayos clínicos con medicamentos. Mosby/Doyma Libros, Madrid, 1994: 45-55.

CARDOSO GARCÍA, Y. Subsistemas de almacenamiento e integración LeukoCIM para el almacén de datos de los Ensayos Clínicos del Centro de Inmunología Molecular. La Habana: Trabajo de Diploma para optar por el título de Ingeniero en Ciencias Informáticas, 2013.

CASTILLO MOREIRA, M.M. Evaluación empírica del Acoplamiento de Algoritmos de Minería de Datos a un Sistema Gestor de Base de Datos. Chile: Tesis presentada en opción al título de Máster en Ingeniería Informática, 2014.

FILIBERTO, Y.; BELLO, R.; CABALLERO, Y. Algoritmo para el aprendizaje de reglas de clasificación basado en la teoría de los conjuntos aproximados extendida. No. 169, Medellín: Revista DYNA, 2011, Vol. 78, 2011.

GASPAROVICA, M.; ALEKSEJEVA, L. Using Fuzzy Algorithms for Modular Rules. Scientific Journal of Riga Technical University Computer Science. Information Technology and Management Science. 2010. págs. 94-98.

IBM. Big Data and analytics platform. Disponible en: http://www-01.ibm. com/software/data/bigdata, 2012.

MINISTERIO DE SALUD PÚBLICA. Centro para el control estatal de la calidad de los medicamentos. Regulación No. 45-2007. [En línea] 2009. [Citado el: 4 de marzo de 2015] http://www.bvv.sld.cu.

MOLINA LÓPEZ, J. M.; GARCÍA HERRERO, J. Técnicas de análisis de datos. 2006.

POSTGRESQL. [En línea] 12 de septiembre de 2011. [Citado el: 13 de septiembre de 2015.] http://www.postgresql.org/about/press/presskit91/es/.

RODRÍGUEZ SUÁREZ, Y. Herramientas de Minería de Datos. La Habana : Revista Cubana de Ciencias Informáticas, 2009.

ROBLES ARANDA, Y. Propuesta de integración de las técnicas de minería de datos Árboles de decisión y reglas de inducción al Sistema Gestor de Base de Datos. La Habana : UCIENCIA, 2012.

ROBLES ARANDA, Y. Algoritmos de minería de datos: Árboles de decisión y reglas de inducción integrados a PostgreSQL. La Habana: Tesis presentada en opción al título de Máster en Informática Aplicada, 2012.

RUIZ, OMAR. Aplicación de minería de datos para detección de patrones en investigaciones biotecnológicas. s.l. : Revista Espol Ciencia, 2008.

USAMA FAYYAD, G. Data Mining and Knowledge Discovery in Databases: An overview, Communications of ACM. [En línea] 1996. [Citado el: 14 de abril de 2015.] http://dl.acm.org/citation.cfm?id=240464.

WITTEN, IH AND FRANK, E. "Data Mining: Practical Machine Learning Tools and Techniques", 2nd Edition. Morgan Kaufmann, 2005.