Optimización de los Diagnósticos Médicos Mediante Técnicas de Minería de Datos

Autoras: Marcela Baukloh Coronil[1]; Carina Yoshimura Kumano[2]

Resumen

El presente trabajo tuvo como objetivo definir patrones respecto a los pacientes bajo condiciones similares, para la optimización de los diagnósticos médicos mediante técnicas de minería de datos, específicamente técnicas de clasificación aplicadas a los registros la base de datos del Centro Latinoamericano de Perinatología (CLAP), que posee el historial de embarazadas que asisten a sus controles prenatales en el Hospital Regional de Encarnación (HRE), comprendidas entre los años 2009 y 2014.Durante la investigación se analizaron las posibles causas de hipertensión arterial, parto prematuro y rotura prematura de membrana en las embarazadas. Los clasificadores indicaron un alto porcentaje de precisión en los resultados obtenidos, por ejemplo, un clasificador indicó 95% de exactitud al determinar como principal causa de la hipertensión arterial (HTA) inducida en el embarazo al antecedente de HTA de la paciente.Considerando el alto porcentaje de exactitud de los clasificadores se concluye que los patrones definidos por los algoritmos de clasificación son válidos para el diagnóstico médico de control prenatal, y por medio de estos patrones se confirman varias teorías médicas como la importancia de las consultas prenatales, la incidencia de los antecedentes personales o familiares, entre otros.

Palabras Claves: Minería de Datos, Clasificación, Patrones, Diagnóstico Médico, Embarazo.

Abstract

The objective of this paper was to define patterns of patients under similar conditions, for the optimization of medical diagnostics using data mining techniques, specifically classification techniques applied to Latin American Center of Perinatology database, which have the history of pregnant attending prenatal checkups in the Encarnación Regional Hospital, between 2009 and 2014. During the implementation of the project, it was carried out the analyses of possible causes of high blood pressure; premature birth and premature membrane rupture in pregnant. Classifiers indicated a high percentage of accuracy in the results, for example, a classifier indicated 95% accuracy determining the main cause of high blood pressure induced by pregnancy the hypertension historic of the patient. Considering the high percentage of accuracy of the classifiers, the conclusion is that the patterns defined by the classification algorithms are valid for the medical diagnosis of prenatal care, and through these patterns, some medical theories are confirmed, such as the importance of prenatal consultations, the incidence of personal or family history, and others.

Keywords: Data Mining, Classification, Patterns, Medical Diagnosis, Pregnancy

[1]Profesora Investigadora de la UNI e-mail: mbaukloh@uni.edu.py
[2]e-mail:carina@datexpy.com

Recibido: 29/07/2016 Aceptado: 10/11/2016

Introducción

La toma de decisiones en el diagnóstico médico es una de las tareas más importantes de los profesionales médicos, y la tasa de diagnósticos incorrectos en la práctica clínica se ha estimado en 150 de 1000 pacientes (según Health Grades Patient Safety in American Hospitals Study, 2011). Humanamente es imposible evitar cometer errores y esto causa daño, gasto y muerte en los centros de salud, especialmente en diagnósticos serios como apendicitis, infección y cáncer (Lugo, Maldonado, & Murata, 2014).

Las técnicas de minería de datos pueden facilitar enormemente el proceso del diagnóstico médico, considerando que los resultados no dependerán de la experiencia previa ni del estado en el que se encuentra el médico en ese momento. En este caso, las técnicas de clasificación se utilizan para determinar patrones genéricos que caracterizan a los pacientes que acuden a consulta y ayuda a predecir una enfermedad; es por ello que resulta importante aplicar estos algoritmos en nuestro estudio sobre optimización de diagnósticos médicos.

En esta investigación se analizó la eficiencia de los algoritmos de clasificación para determinar patrones patológicos en embarazadas, como ser hipertensión inducida por el embarazo, partos prematuros, rotura prematura de membrana, anemia durante el embarazo, entre otros.

Materiales y Métodos

Luego se dio inicio al proceso de minería de datos, el cual se llevó a cabo utilizando la metodología "Proceso estándar aplicado a la industria para la minería de datos" (CRISP-DM). Como primer paso, se analizó la base de datos con las herramientas Microsoft Access y Microsoft Excel, llegando a seleccionar las tablas nivel 01 y nivel 05, las cuales eran adecuadas para aplicar técnicas de minería de datos. La tabla nivel 01 cuenta con 235 atributos, un ejemplo de estos datos se puede observar en la Tabla 1, donde se presenta la denominación de cada atributo y los posibles valores que puede tomar cada uno de estos.

Tabla 1
Fragmento de la tabla Nivel 01

Denominación	Posibles Valores		
edad_riesgo	x		
etnia	blanca	indígena	mestiza
diabetes_familiar	no	si	
hipertension_familiar	no	si	
preeclampsia_familiar	no	si	
eclampsia_familiar	no	si	
hipertension_personal	no	si	
grupo_sanguineo	texto		
toxo_igg_menor_veinte_sem	-	+	no se hizo
toxo_igg_mayor_veinte_sem	-	+	no se hizo
bacteriuria_menor_veinte_sem	normal	anormal	no se hizo

Por su parte, la tabla nivel 05 cuenta con 28 atributos, un ejemplo de estos datos se puede observar en la Tabla 2, donde también se presenta la denominación de cada atributo y los posibles valores que pueden tomar estos.

Tabla 2
Fragmento de la tabla Nivel 05

Denominación	Posibles Valores			
embarazo_ectopico	numérico			
fuma_act_1	no	si		
drogas_1	no	si		
alcohol_1	no	si		
sif_no_trepo_valor1	numérico			
sif_trepo1	-	+	se desconoce	no corresponde

Para la comprensión de estos datos se llevaron a cabo varias entrevistas con los profesionales del área, donde se identificaron cada una las variables de la base de datos, sus respectivos valores y las posibles relaciones entre ellas.

Posteriormente, se llevó a cabo el pre-procesamiento de los datos, por medio de la herramienta Navicat, donde fue necesario importar la base de datos e ir aplicando distintas consultas con el propósito de tener datos íntegros y adaptados para su explotación.

Una vez finalizado el pre-procesamiento se realizaron pruebas en la herramienta RapidMiner con dos algoritmos de clasificación, las Redes Bayesianas y los Árboles de Decisión, a partir de estas pruebas se analizaron cuáles de estos ofrecían mejores resultados para la explotación de información con la base de datos CLAP,

adaptándose mejor a los objetivos de la investigación, el algoritmo Árbol de Decisión.

A partir de los resultados obtenidos, se identificaron varios patrones que podrían ayudar a optimizar los diagnósticos médicos en embarazadas. Se obtuvieron algunos resultados novedosos y que efectivamente serían difíciles de determinar bajo la manipulación de los datos con métodos convencionales.

Resultados

El objetivo del estudio fue determinar patrones y comportamientos de enfermedades o patologías para optimizar los diagnósticos médicos utilizando técnicas de minería de datos, en este caso Árboles de Decisión.

Se obtuvo un buen comportamiento del clasificador, alcanzando una precisión mayor al 70% observado por medio de la matriz de confusión; además con ayuda de los profesionales médicos fue posible validar los patrones obtenidos.

A partir de lo mencionado, se pueden considerar como válidos los siguientes patrones en las embarazadas:

• Las pacientes que realizan menos de 5 consultas prenatales tienen 11% más de probabilidad de experimentar un parto prematuro, según el clasificador que presenta una exactitud del 84%. Además, tienden a padecer niveles bajos de hemoglobina y no realizan todos los estudios requeridos por el médico. A partir de esto se confirma que la cantidad de consultas prenatales es una variable primordial para el buen desarrollo del embarazo.

• La hipertensión inducida por el embarazo es una de las patologías que ponen en riesgo tanto la vida de la madre como la del feto, en este estudio se pudo observar que efectivamente existe mayor probabilidad de que una paciente sufra HTA inducida si tiene antecedentes de hipertensión personal o familiar. Finalmente, las embarazadas que tienen antecedentes de HTA tienen 29% más de probabilidad de sufrir HTA inducida por el embarazo, con una precisión del 95% según se observa en la matriz de confusión de la Figura 1.

Figura 1.
Matriz de confusión

Accuracy:95.06% +/- 0.68(mikro: 95.06%)			
	True No	true SI	class precisión
pred. NO	14716	512	96.64%
Pred. SI	268	279	51.01%
Class recall	98.21%	35.27%	

• Se observaron patrones donde el 50% de las pacientes que sufren HTA inducida por el embarazo también tienen hemoglobina baja, antes o después de las 20 semanas, según se indica en el clasificador que cuenta con un 94% de exactitud.

• Las embarazadas que realizan menor cantidad de consultas prenatales, que sufren de infección urinaria y que no cumplen con los exámenes solicitados por el médico tienen mayor riesgo de experimentar un parto prematuro, esto revela el clasificador con una precisión del 84%.

• Se observaron que el 64% de los casos de abortos y muertes fetales se dieron en pacientes que no se realizaron los exámenes de estreptococo b y toxoplasmosis igg antes de la semana 20 de gestación, con una precisión del 74% en el clasificador.

• En un clasificador con 94% de precisión se observó que existe 6% mayor probabilidad de que una paciente con factor RH Positivo sufra HTA Inducida en comparación con las embarazadas de factor RH Negativo.

Discusión

Mediante este estudio fue posible afirmar varias teorías médicas, como ser, que la causa principal de HTA inducida en el embarazo es el antecedente de HTA personal o familiar; y que

aquellas pacientes que asisten a una menor cantidad consultas prenatales o sufren infección urinaria tienen mayor probabilidad de experimentar parto prematuro.

Por otra parte, se pusieron a consideración algunos resultados que no coinciden con las teorías médicas, como, por ejemplo, que el 50% de las pacientes con HTA inducida por el embarazo sufren de hemoglobina baja, o que un mayor porcentaje de embarazadas con tipo de sangre RH positivo tenga HTA inducida, sin embargo, son puntos que deberían ser estudiados de forma específica y a mayor profundidad para finalmente afirmar o descartar las nuevas informaciones.

Conclusión

El resultado del presente trabajo, demuestra que para determinar patrones médicos y optimizar los diagnósticos, el algoritmo Decision Tree (Árbol de Decisión) ofrece un mejor comportamiento en comparación al algoritmo Redes Bayesianas, considerando la estructura de los resultados expuestos por el algoritmo Árbol de Decisión, el cual ofrece resultados en forma de árbol y de reglas, que son realmente útiles y comprensibles al momento de buscar explicaciones a ciertos comportamientos en los diagnósticos médicos.

Mediante las reglas obtenidas se determinan patrones válidos para el diagnóstico de hipertensión inducida por el embarazo, rotura prematura de membrana, parto pretermino y muerte fetal. Además, por medio de estos patrones se comprueban algunas teorías médicas como la importancia de las consultas prenatales o la realización de todos los estudios prenatales solicitados para un control riguroso de la madre y el feto, y así evitar un parto pretermino o la muerte fetal.

Luego de analizar el nivel de precisión de los clasificadores y hacer un control cruzado con los profesionales médicos se comprueba que los algoritmos de clasificación son aptos para determinar patrones que expliquen el comportamiento de las patologías y así optimizar los diagnósticos médicos, brindando un apoyo eficaz a los profesionales de la salud, siempre que los datos proveídos sean íntegros y confiables.

Bibliografía:

- Franco-Arcega, A., Carrasco-Ochoa, J. A., Sánchez-Díaz, G., & Martínez-Trinidad, J. F. (2013). Decision Tree based Classifiers for Large Datasets.Computación y Sistemas, 17(1), 95-102.
Britos, P., Hossian, A., García Martínez, R., & Sierra, E. (2005). Minería de Datos Basada en Sistemas Inteligentes. 876 páginas. Editorial Nueva Librería. ISBN 987-1104-30-8.

- Chaurasia, V., & Pal, S. (2014). Data Mining Approach to Detect Heart Diseases. International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol, 2, 56-66. Dávila, F., & Sánchez, Y. (2012). Técnicas de minería de datos aplicadas al diagnóstico de entidades clínicas. Revista Cubana de Informática Médica.

- Dávila Hernández, F., & Sánchez Corales, Y. (2012). Técnicas de minería de datos aplicadas al diagnóstico de entidades clínicas. Revista Cubana de Informática Médica, 4(2), 174-183.Jiawei Han; Jian Pei; Yiwen Yin; Runying Mao. (2004). Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. 8(1).

- Pinho Lucas, J. (2010). Métodos de clasificación basados en asociación aplicados a sistemas de recomendación.

- Lugo-Reyes, S. O. (2014). Inteligencia artificial para asistir el diagnóstico clínico en medicina. Órgano oficial del Colegio Mexicano de Inmunología Clínica y Alergia, AC y de la Sociedad Latinoamericana de Alergia, Asma e Inmunología, 61, 110-120.

- Quesada, Y. A., Pérez, D. W., & Suárez, A. R. (2012, June). Minería de Datos aplicada a la Gestión Hospitalaria. In V Simposio de Ingeniería Industrial y Afines.

- Hofmann, M., & Klinkenberg, R. (Eds.). (2013). RapidMiner: Data mining use cases and business analytics applications. CRC Press.