ISSN 0798 1015

REVISTA ESPACIOS

HOME | Revista ESPACIOS | ÍNDICES / Index | A LOS AUTORES / To the AUTHORS

Vol. 39 (Nº 28) Año 2018 • Pág. 16

El diagnóstico de enfermedades desde el Análisis Inteligente de los Datos

The diagnosis of diseases from the intelligent analysis of data

Maikel LEYVA VÁZQUEZ 1; Neilys GONZÁLEZ BENÍTEZ 2; Jesús HECHAVARRÍA HERNÁNDEZ 3; Yasnalla RIVERO PEÑA 4; Jorge Elías DAHER NADER 5

Recibido: 16/02/2018 • Aprobado: 25/03/2018

Contenido

RESUMEN:
El desarrollo acelerado de la tecnología ha permitido el almacenamiento y tratamiento de grandes volúmenes de información compuesta de diferentes tipos de datos, los cuales no siempre son tan precisos y completos como se hace necesario. Por lo anterior el pre procesamiento de datos constituye un paso previo para obtener datos de calidad y a partir de ellos ejecutar diagnósticos de enfermedades utilizando técnicas de IA, en particular las redes bayesianas.
Palabras-Clave: Diagnóstico de enfermedades, Análisis Inteligente de los Datos, Pre procesamiento de datos

ABSTRACT:
Medicine faces the challenge of acquiring, analyzing and applying knowledge to solve complex clinical problems. The accelerated development of technology has allowed the storage and processing of large volumes of information composed of different types of data, which are not always as accurate and complete as necessary. Therefore, the pre-processing of data is a step prior to obtaining quality data and from them perform disease diagnoses using AI techniques, in particular the Bayesian networks.
Keywords: Diagnosis of diseases, Intelligent Data Analysis, Pre-processing of data.

1. Introducción

Para el diagnóstico de enfermedades desde los datos almacenados en bases de datos, Cuba cuenta con herramientas para el procesamiento de la información respecto a las enfermedades más recurrentes, obteniéndose desde ellos una estadística del comportamiento de dichas enfermedades.

Los datos que se registran con frecuencia son utilizados para encontrar la prevalencia, morbilidad y mortalidad sobre cualquier enfermedad de interés, resultados que se obtiene con técnicas estadísticas, en particular para los estudios relacionados con la búsqueda de factores de riesgo, donde se utiliza la estadística descriptiva y los modelos basados en regresiones para descubrir conocimiento sobre una población. Sin embargo, la búsqueda de relaciones relevantes entre los datos médicos no es buena ya que no se emplean técnicas que posibilitan encontrar nuevos conocimientos a partir de los datos que se tienen.

Se maneja un elevado número de datos, relacionados con las informaciones relativas al comportamiento de una determinada enfermedad en una población, y que son utilizados en este contexto para predecir el comportamiento de las enfermedades.

En Bernal (2014) se reportan aplicaciones relacionadas con diagnósticos de enfermedades, donde previamente se ejecutó un proceso de Minería de Datos, con el fin de obtener datos de calidad y desde los datos, ejecutar procesos de diagnósticos de enfermedades utilizando técnicas de IA. Esas aplicaciones demostraron resultados favorables en diversas enfermedades de salud en humanos

Las técnicas de Inteligencia Artificial (IA), a partir de la década del 50, han sido utilizadas y aplicadas en diversas áreas comunes al desempeño cotidiano de las personas, ayudando, reemplazando, simulando las acciones o decisiones tomadas por individuos con ciertas características en particular. Como una de sus principales áreas de aplicación se puede destacar particularmente la medicina, dado que la Inteligencia Artificial logró su mayor impacto inicial en ella a través de los sistemas expertos y específicamente por medio de los sistemas de diagnóstico (Bernal, 2014).

González (2017), refiere que despierta un marcado interés y utilidad el empleo de las técnicas de IA, debido a las posibilidades que presentan para involucrarse en situaciones donde se requiere de un gran acervo de conocimientos, el veloz procesamiento de datos y la toma efectiva de decisiones, por lo que los procesos de diagnóstico de enfermedades, son más certeros al tratar los mismos con técnicas de Inteligencia Artificial desde datos consistentes.

Refiere, además, que en el área médica existe necesidad de encontrar nuevas soluciones para el procesamiento de la información, análisis desde los datos, diagnósticos de enfermedades con un mayor grado de certeza. Posibilitando la obtención de resultados favorables para apoyar la toma de decisiones.

Específica, González (2017) que después de tratados los datos se evitan obstáculos que en ocasiones se presentan, debido a la coincidencia de uno o varios factores; como la escasez de expertos en el área médica o la abundancia/ausencia y dispersión de la información disponible. Influyendo, tales deficiencias sobre una variable crítica perteneciente al diagnóstico, el tiempo transcurrido desde la observación de los síntomas, hasta la determinación de los agentes patógenos, responsables de la presencia de una determinada enfermedad.

Tomando en consideración la importancia del significativo volumen de datos que se generan, relacionados con síntomas clínicos de diversas enfermedades, se requiere un Análisis Inteligente de los Datos con el objetivo de poder tratar los datos con técnica que no produzcan pérdida de información relevante y que puedan afectar de forma directa la calidad de los resultados obtenidos. Por tal motivo se propone como principal objetivo ejecutar el pre procesamiento de datos para obtener datos de calidad y ejecutar diagnóstico de enfermedades desde ellos aplicando técnicas capaces de extraer la máxima información, y a su vez se pueda obtener resultados lo más certeros posible.

---

2. Materiales y métodos

Para ejecutar el diagnóstico de enfermedades desde los datos con mayor certeza se tuvo en consideración que en el área médica los datos provienen de fuente de datos diferentes, por lo que se hace necesario el pre procesamiento de datos para eliminar los datos que causen contratiempo en los resultados esperados, en aras de garantizar su disponibilidad, completitud y fidelidad.

El pre procesamiento de datos que se ejecuta es a través de la tarea de limpieza de datos, para luego incorporar los datos limpios a una Base de Datos que se crea previamente y a partir de ella ejecutar diagnóstico de enfermedades utilizando técnicas de IA. Esta tarea se realiza de forma automática, utilizando el algoritmo K–Means.

El algoritmo K-Means tienen como base la optimización de una función criterio, donde en el presente trabajo, se denomina F, el valor de esta función depende de las particiones del conjunto de datos $\{C_1, \ldots, C_k\}$

$$F: P_k(X) \rightarrow \mathbb{R} \qquad (1)$$

Donde:

$P_k(X)$, son las particiones del conjunto de datos $X = \{x_1, \ldots, x_n\}$ en K grupos no vacíos. x_i, es un vector $n\text{-}dimensional$ (objeto) del conjunto de datos X.

El algoritmo K-Means converge a un mínimo local, utilizando la función criterio F, de la sumatoria de las distancias $L2$ entre cada objeto y su centroide más cercano. A este criterio normalmente se le denomina error cuadrático y se obtiene a través de la expresión 2.

$$F(\{C1, \ldots, CK\}) = \sum_{i=1}^{K} \sum_{j=1}^{p_i} \| x_{i_j} - \bar{C}i \| \qquad (2)$$

Donde:

K es el número de grupos, pi es el número de objetos del grupo i, x_{i_j} es el $j\text{-}ésimo$ objeto del $i\text{-}ésimo$ grupo y $\bar{C}i$ es el centroide del $i\text{-}ésimo$ grupo el cual es calculado a través de la expresión 3.

$$\bar{C}i = \frac{1}{p_i} \sum_{j=1}^{p_i} X_{i_j}, i = 1, \ldots K \qquad (3)$$

El conjunto de pasos lógicos del algoritmo K-Means es el que se presenta a continuación:

Paso 1. Selecciona los K centroides iniciales $\{C_i, \ldots, C_k\}$.

Paso 2. Asigna los objetos x_i del conjunto de datos X, a su centroide más cercano.

Paso 3. Recalcula los nuevos centros, regresa al paso 2, hasta que el algoritmo converge.

El algoritmo se inicia seleccionando o calculando los centroides iniciales, dependiendo del criterio de selección de centroides, posteriormente asigna los objetos a su centroide más cercano, para después recalcular los nuevos centroides esto lo realiza hasta que el algoritmo converja (paso 3).

Pre procesados los datos y aplicada la técnica de limpieza de datos, los datos faltantes se rellenan, utilizando el método de imputación por media condicional. Método que sustituye los valores faltantes de una variable mediante la media de las unidades observadas en esa variable.

El método de imputación de valores faltan tes contribuye a reducir la pérdida de los datos faltantes en la base de datos (Castro, 2014). Las técnicas de imputación se pueden clasificar, en primer lugar, en dos grandes grupos: las técnicas de imputación simples y las de imputación múltiple. En el presente trabajo se utilizan técnicas simples de imputación, ella según refieren (Saransk, 1985, Kalton y Kasprzyk, 1986 y Little y Rubin, 2002) han sido una de las herramientas más conocidas y aceptadas para el tratamiento de la falta de respuesta.

3. Resultados y discusión

El proceso de limpieza de datos se debe a la inconsistencia de los datos almacenados, los cuales poseen ruidos, para ello se utilizó el algoritmo K – Means de forma general, que se basa en el trabajo con datos continuos, para el diagnóstico de enfermedades los datos que se utilizan son binarios, como se muestra en la tabla 1, por tal motivo para la limpieza de este tipo de datos se utiliza el algoritmo K - Modas (análogo a K-Means).

Tabla 1
Síntomas reportados sobre la notificación de enfermedades en las poblaciones. Fuente: Elaboración propia.

Notificación de enfermedad	Síntoma_1	Síntoma_2	Síntoma_3	Síntoma_4	Síntoma_5	Síntoma_6	Síntoma_7	Síntoma_8
Población_1	1	1	0	1	0	0	1	0
Población_2	1	1	1	1	0	1	0	0
Poblaación_3	1	1	1	1	1	1	0	1
Poblaación_4	1	0	0	0	0	1	1	1
Poblaación_5	0	1	0	0	0	1	1	1
Poblaación_6	1	1	0	1	0	0	1	0
Poblaación_7	1	1	1	1	0	1	0	0
Poblaación_8	1	1	1	1	1	1	0	1
Poblaación_9	1	0	0	0	0	1	1	1
Poblaación_10	0	1	0	0	0	1	1	1

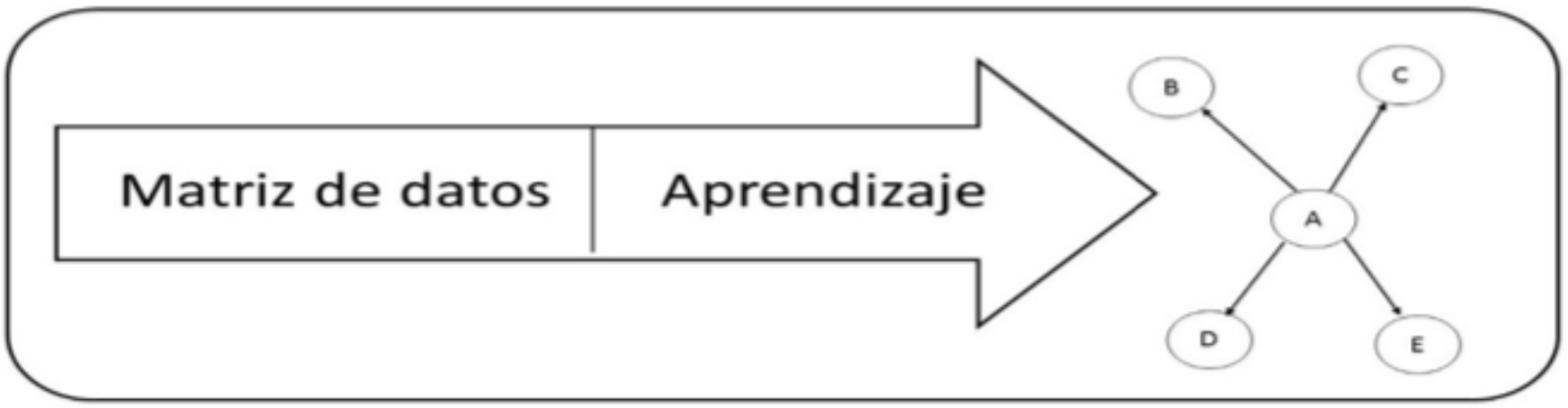
Para ejecutar la limpieza de datos, utilizando el algoritmo K-Means, los datos de entrada que son los que se muestran en la tabla 1, se dividen en dos grupos (K=2), el primer grupo se corresponde con los enfermos y el segundo grupo se corresponde los sanos. La política seguida para seleccionar el primer grupo K= 1, que es el centroide correspondiente a los enfermos, es seleccionar la población que más síntomas presentes posea de los síntomas representativos para la enfermedad que se estudia, para el K =2, correspondiente al segundo centroide, la política seguida para su selección es que las poblaciones sanas no posean ninguno de los síntomas representativos para la enfermedad que se estudia.

La división de grupos favorece el apoyo a la toma de decisiones en el diagnóstico de enfermedades y en consecuencia contribuye a que el tratamiento médico que se aplique sea el más adecuado. Obtenidos los dos grupos se calcula la moda del grupo de la población enferma y con el resultado obtenido se rellenan los valores incompletos, en la variable cantidad de enfermos, variable que mayor imperfección de datos presenta.

Luego de realizar la limpieza de datos, los datos faltantes son imputados por la media condicional debido a las características de los datos, para ellos se calculó el valor medio de la muestra total, lo cual permitió a partir de ese valor obtener los nuevos valores a reemplazar. Este método posee como principal desventaja que distorsiona la distribución de los datos, debido a la concentración de valores en torno a la media. Para obtener mejores resultados al aplicar el método, Cohen (1996) propuso añadir más variabilidad a los valores

imputados usando la variabilidad de los datos muéstrales. Este procedimiento, se realiza tomando el valor de la media con respecto a la muestra total de datos a imputar para primeramente imputar la mitad de los valores faltantes y posteriormente imputar los datos faltantes de la otra mitad, obteniendo finalmente el conjunto de datos consistentes que facilitan ejecutar el diagnóstico de enfermedades desde una matriz de datos como se muestra en la figura 2.

Figura 1
Proceso de ejecución de diagnóstico de enfermedades desde los datos. Fuente: Elaboración propia.



4. Conclusiones

En el presente trabajo se obtuvo una matriz de datos consistente útil para ejecutar diagnóstico de enfermedades desde ellos, se realizó una limpieza de datos y se imputaron los datos faltantes lo que facilito el trabajo con datos certeros, los resultados obtenidos están de acuerdo a los reportados en la literatura para el diagnóstico de enfermedades y son utilizados para apoyar la toma de decisiones. Como trabajos futuros se plantea el empleo de métodos de diagnóstico médico como los mapas cognitivos difusos (Leyva-Vázquez, Santos-Baquerizo, Peña-González, Cevallos-Torres, & Guijarro-Rodríguez, 2016).

Referencias bibliográficas

Bernal, E., A. (2014). Sistema prototipo de entrenamiento pediatra para el proceso de adaptación neonatal, (Tesis de maestría en Ingeniería de Sistemas y Computación), Línea de investigación: Ingeniería del Software, Sistemas Inteligentes, ÁREA: Sistemas Inteligentes.

Castro, M. (2014). Imputación de datos faltantes en un modelo de tiempo de fallo acelerado. (Tesis de fin de Máster en Técnicas Estadísticas). Universidad de Santiago de Compostela, Galicia, España.

Cohen, M.P. (1996). A new approach to imputation. American Statistical Association Proceding of the Section on Survey Research Methods 293 - 298.

Hu, M., Salvucci, S. y Lee, R. (2001). A Study of Imputation Algorithms. Working Paper No. 200117. Washington DC: U.S. Department of Education, National Center for Education Statistics, 2001. 27 Stata Statistical Software.

González, N. (2017). Modelo basado en redes bayesianas para el diagnóstico de la Fasciolosis bovina. (Tesis en opción al grado científico de Doctor en Ciencias Técnicas). Universidad de las Ciencias Informáticas. Ciudad de la Habana, Cuba.

Kalton, G. y Kasprzyk, D. (1986). The treatment of missing survey data. Survey Methodology 12 1-16.

Leyva-Vázquez, M., Santos-Baquerizo, E., Peña-González, M., Cevallos-Torres, L., & Guijarro-Rodríguez, A. (2016, November). The Extended Hierarchical Linguistic Model in Fuzzy Cognitive Maps. In International Conference on Technologies and Innovation (pp. 39-50). Springer, Cham.

Little, R.J.A. y Rubin, D.B. (2002). Statistical analysis with missing data. 2nd edition. New York: John Wiley & Sons, Inc.

Muñoz, J.F., (2009). Métodos de imputación para el tratamiento de datos faltantes: aplicación mediante R/Splus. REVISTA DE métodos CUANTITATIVOS PARA LA economía Y LA EMPRESA (7). 3–30. Junio de 2009. ISSN: 1886-516X. D.L: SE-2927-06. URL: http://www.upo.es/RevMetCuant/art25.pdf

Sedransk, J. (1985). The objetive and practice of imputation. In Proc. First Annual Res. Conf., Washington, D.C.: Bureau of the Cencus. 445 - 452.

---

1. PhD. En Ciencias Técnicas, , MSc. En Bioinformática, Facultad de Ciencias Ciencias Médicas, Universidad de Guayaquil, Guayaquil Ecuador. mleyvaz@gmail.com

2. PhD. En ciencias Técnicas. Centro Meteorológico Provincial de Pinar del Río, Cuba, neilysgonzalezbenítez@gmail.com

3. PhD. En Ciencias Técnicas , Universidad de Guayaquil, Facultad de Arquitectura y Urbanismo, Ecuador, jesus.hechavarria@cu.ucsg.edu.ec

4. Master en Ciencias, Universidad de Holguín, Cuba, yasnidf@gmail.com

5. Dr y Msc. , Facultad de Ciencias Médicas, Facultad de Ciencias Médicas, Universidad de Guayaquil, jdaher_nader@yahoo.es

---

Revista ESPACIOS. ISSN 0798 1015
Vol. 39 (Nº 28) Año 2018

[Índice]

[En caso de encontrar un error en esta página notificar a webmaster]