Propuesta de Refinamiento de un Algoritmo de Minería de Datos para Detección de Pacientes

Omar Fernando García-Mora, Federico Miguel Cirett-Galán, Raquel Torres-Peralta
Departamento de Ingeniería Industrial
Universidad de Sonora
Hermosillo, México
a218230084@unison.mx, federico.cirett@unison.mx, raquel.torres@unison.mx

***Resumen*— En el presente artículo se propone una estrategia para el refinamiento de un algoritmo de agrupamiento automático perteneciente al área de medicina preventiva de una Institución de Salud Pública para mejorar la detección de pacientes diabéticos. En particular, se presenta una estrategia para aumentar la generación de conocimiento a través del uso de técnicas de minería de datos, en la que se propone agregar variables que permitan describir mejor al grupo de pacientes para mejorar la precisión del algoritmo.**

Palabras claves: Minería de Datos, Generación de Conocimiento, K-means, Medicina Preventiva

I. INTRODUCCIÓN

En la última década ha habido un uso cada vez mayor de técnicas de minería de datos en el sector salud para descubrir tendencias o patrones útiles que se utilizan en el diagnóstico y la toma de decisiones [1]. Esto, debido a que las instituciones de salud poseen una gran cantidad de datos generados de los procesos de atención médica, los cuales pueden convertirse en conocimiento y tener un propósito útil [2]. Los algoritmos de minería de datos, cuando se usan apropiadamente, son capaces de mejorar la calidad de la predicción, el diagnóstico y la clasificación de enfermedades [1].

En el presente artículo se muestra una propuesta de estrategia para el refinamiento de un algoritmo de minería de datos enfocado en medicina preventiva para mejorar la atención de pacientes diabéticos. La aplicación de dicha estrategia busca crear un impacto positivo en la calidad de vida de la población, ya que está enfocado en la segunda mayor causa de muerte a nivel nacional y tercera a nivel estatal, como lo es la diabetes mellitus. Además, repercute en la detección de pacientes que ya sufren de la enfermedad pero no han sido detectados, el cual corresponde al 50% de las personas que padecen de dicho padecimiento, según la Federación Internacional de Diabetes.

Al estar enfocado en medicina preventiva el impacto se puede reflejar en ahorros para la institución ya que se pueden llegar a evitar consultas, tratamientos y medicamentos en derechohabientes.

Primeramente se presenta el marco teórico, el cual sirve como base para un mejor entendimiento de los conceptos básico de la problemática a resolver. Después se describe la problemática actual, seguido por la propuesta de solución a implementar, los resultados esperados al aplicar dicha propuesta, y por último las conclusiones.

II. MARCO TEÓRICO

Los orígenes de la Minería de Datos (MD) se remontan a partir de los años 60, cuando se basaba simplemente en el procesamiento de archivos [3]. El crecimiento explosivo de las bases de datos ha creado la necesidad de desarrollar tecnologías que utilicen la información y el conocimiento de manera inteligente. Por lo tanto, la MD se ha convertido en un área de investigación cada vez más importante y es uno de los componentes principales en el proceso de Descubrimiento de Conocimiento en Bases de Datos [4].

Milley [5] define la MD como el proceso de selección de datos y modelos de exploración y construcción que utilizan vastos almacenes de datos para descubrir patrones previamente desconocidos. En adición, Durairaj y Ranjani [3] mencionan que los algoritmos de MD aplicados en la industria de la salud desempeñan un papel importante en la predicción y el diagnóstico de las enfermedades.

Los algoritmos de MD se clasifican en dos categorías: modelo descriptivo, o aprendizaje no supervisado, y modelo predictivo, o aprendizaje supervisado [6]. La minería de datos descriptiva agrupa los datos al medir la similitud entre objetos, o registros, y descubre patrones o relaciones desconocidos en los datos para que los usuarios puedan comprender fácilmente una gran cantidad de datos. Este tipo de modelos incluye técnicas como agrupamiento, asociación, resumen y descubrimiento de secuencias. Por su parte, La minería de datos de predicción deduce reglas de predicción a partir de datos de entrenamiento y aplica las reglas a datos no predichos o no clasificados. Los modelos predictivos incluyen técnicas como clasificación, regresión, análisis de series de tiempo y predicción [7].

Según Patil [8], las técnicas de MD, como asociación, clasificación y agrupamiento, son utilizadas por la organización de atención médica para aumentar su capacidad para construir conclusiones apropiadas con respecto a la salud del paciente a partir de datos y cifras sin procesar.

La asociación tiene un gran impacto en la industria del cuidado de la salud para descubrir las relaciones entre las enfermedades, el estado de la salud humana y los síntomas de la enfermedad [9].

Por su parte, la clasificación comprende de dos pasos: 1) Entrenamiento y 2) Pruebas. El entrenamiento construye un

modelo de clasificación, el cual consiste en reglas de clasificación, mediante el análisis de datos de entrenamiento que contienen etiquetas de clase. El segundo paso, la prueba, examina un clasificador, utilizando datos de prueba, para determinar la precisión o capacidad para clasificar registros desconocidos para la predicción [10].

El agrupamiento divide los datos en función de las similitudes que tiene. Los algoritmos de agrupación descubren colecciones de datos de manera que los objetos en la misma agrupación son más idénticos entre sí que otros grupos [11]. De acuerdo con Haraty, Dimishkieh y Masud [12], el algoritmo de agrupamiento K-means es uno de los métodos de agrupación de datos más utilizados, el cual tiene como función la división de un conjunto de datos de "n" observaciones en "k" grupos, en donde cada dato observado corresponde al grupo "k" cuyo centroide es más cercano.

Existen varios estudios relacionados al tema, como el de Hartono et al. [13], donde se propone un enfoque para optimizar un algoritmo de agrupamiento K-means para resolver problemáticas de precisión de predicción y sesgo en la toma de decisiones provocado por un desequilibrio de clases.

Además, existen trabajos que han utilizado la misma base de datos que se utilizará para desarrollar el presente estudio, como el de Sanez [14], quien implementó una estrategia basada en análisis de datos para analizar registros de consultas para detectar grupos a los cuales dirigir campañas de medicina preventiva en obesidad y diabetes. El presente se enfoca específicamente en mejorar las etapas de segmentación (etapa II) y de uso de MD (etapa III) de dicha estrategia.

III. Entorno del Problema

El trabajo se desarrollará en una institución de salud pública que presta servicios de seguridad social a un aproximado de 180,000 derechohabientes.

La institución cuenta con una base de datos con el registro de consultas a médico general, la cual se ha ido recabando desde hace varios años. En este trabajo se propone explotar estos registros para ampliar el conocimiento del estado de salud de los derechohabientes y, en específico, optimizar el modelo actual de medicina preventiva en pacientes diabéticos y/o pacientes con riesgo de padecer diabetes.

La metodología utilizada para el programa actual trabaja con el algoritmo K-means, esta decisión fue basada en que demanda poco tiempo para poder ser aplicado en comparación con otros algoritmos, por lo que no se estudiaron resultados con otros algoritmos, pudiendo existir otro más eficiente respecto a tiempo y precisión.

El algoritmo que actualmente está en uso tiene un amplio rango de mejora en la forma en que describe a la población para realizar las predicciones; dicho modelo basa sus decisiones en edad, sexo y herencia. Sin embargo, la mejora del algoritmo es necesaria ya que hay otros factores que pueden ser de utilidad para hacer el modelo más robusto, esto con el objetivo de describir de mejor manera a los diferentes segmentos de población y mejorar la predicción. Lo anterior agregando nuevas variables derivadas del análisis de la información que se encuentra en la base de datos de la institución, como lo son las enfermedades relacionadas antes y después de padecer diabetes.

Según el modelo actual, la edad es un factor que influye en sufrir padecimientos de diabetes, siendo las personas mayores quienes tienen más riesgo de padecer este tipo de enfermedad. Esto provoca que existan pacientes de edad avanzada que no cuentan con padres y madres afiliados, por lo que no se puede hacer un análisis de la herencia de estos.

Al momento, la forma en la que están compuestos los diferentes grupos, en cuanto a integrantes, muestra un desequilibrio que puede estar ocasionando problemáticas de precisión en la predicción, además de sesgo en el proceso de tomas de decisiones. Esto, debido a que existe una diferencia de 535 integrantes entre el grupo con mayor cantidad de integrantes comparado con el grupo minoritario.

IV. Propuesta de Solución

La metodología propuesta en el presente trabajo se compone de cinco etapas, las cuales se presentan en la Fig. 1.

Figura 1. Propuesta de estrategia para refinamiento de algoritmo

A continuación se describe cada una de las etapas:

- Análisis del problema: Fase inicial que implica conocer entorno y situación actual del algoritmo perteneciente al programa de medicina preventiva
- Análisis de datos: Esta fase incluye la familiarización con los datos, identificar las características de la base de datos y descubrir las relaciones más evidentes para establecer las primeras hipótesis de relaciones entre diabetes mellitus y otras enfermedades.
- Preparación de datos: En esta etapa se llevará a cabo la limpieza de los datos, su integración, transformación y reducción de estos.
- Modelado: Se seleccionarán y aplicarán técnicas de minería de datos para definir variables a agregar al algoritmo.
- Evaluación: Una vez mejorado el algoritmo se debe evaluar el rendimiento del mismo con las nuevas variables que lo componen contra el algoritmo actual.

V. Resultados Esperados

Se espera que derivado del refinamiento del algoritmo, el programa de medicina preventiva sea capaz de detectar con mayor precisión a pacientes diabéticos que no hayan sido diagnosticados.

La institución podrá verse beneficiada en ahorros en citas médicas, medicamentos y tratamientos derivados de atender a pacientes en la fase prediabética, en lugar de atender a al paciente una vez que éste padece dicha enfermedad.

Además, los derechohabientes también se verán favorecidos, ya que se mejorará la prevención y detección de

una de las principales enfermedades de causa de muerte a nivel nacional y estatal, como lo es la diabetes.

VI. CONCLUSIONES

La aplicación de la MD en el sector salud se ha vuelto un elemento necesario para poder brindar un mejor servicio médico a los derechohabientes, por consecuencia, la mejora de un algoritmo de MD se vuelve una parte esencial en la mejora continua de los procesos de medicina preventiva. En este caso, al refinar el algoritmo e incluir nuevas variables, permitirá crear mejores conclusiones con la misma información que hasta el momento se tiene y que se recolecta día a día en las consultas médicas, pero que se dejan fuera en el análisis que realiza el algoritmo actual.

Se podrá contar con un algoritmo más robusto, el cual permitirá al programa de medicina preventiva mejorar en su proceso de detección de pacientes con riesgo de padecer diabetes, así como también podrá atacar uno de los principales problemas dicho padecimiento, el cual es que la mitad de las personas que padecen esta enfermedad no son diagnosticadas.

REFERENCIAS

[1] Z. S. Daliri, “Data Mining for Health Care Industry: A Practical Machine Learning Tool”. Int. Res. J. Multidiscip. Stud., vol. 3, no. 4, pp. 45–51, 2017.

[2] N. A. Setiwan, P. A. Venkatachalam y A. F. M. Hani, “Missing Attribute Value Prediction Based on Artificial Neural Network and Rough Set Theory”. 2008 Int. Conf. Biomed. Eng. Informatics, vol. 1, pp. 306–310, 2008.

[3] M. Durairaj y M., Ranjani, “Data Mining Applications in Healthcare: A Study”. Int. J. Sci. Technol. Res., vol. 2, no. 10, pp. 29–35, 2013.

[4] P. Mahindrakar y M. Hanumanthappa, “Data Mining in Healthcare: A Survey of Techniques and Algorithms with its Limitations and Challenges”. Int. J. Eng. Res. Appl., vol. 03, no. 06, pp. 937–941, 2013.

[5] A. Milley, “Healthcare and data mining”. Health Manag. Technol., vol. 21, no. 8, pp. 44–47, 2000.

[6] P. P. Sondwale, “Overview of Predictive and Descriptive Data Mining Techniques”. Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 5, no. 4, pp. 262–265, 2015.

[7] M. Dunham, “Data mining—Introductory and advanced topics”. New Jersey, 2003.

[8] S. L. Patil, “Survey of Data Mining Techniques in Healthcare”. Int. Res. J. Innov. Eng., vol. 01, no. 9, pp. 1–3, 2015.

[9] S. Patel y H. Patel, “Survey of Data Mining Techniques Used in Healthcare Domain”. Int. J. Inf. Sci. Tech., vol. 6, no. 1, pp. 9–10, 2016.

[10] I. Yoo et al. “Data mining in healthcare and biomedicine: A survey of the literature”. J. Med. Syst., vol. 36, no. 4, pp. 2431–2448, 2012.

[11] K. Sharmila y S.A. Vethamanickam, “Survey on Data Mining Algorithm and Its Application in Healthcare Sector Using Hadoop Platform”. Int. J. Emerg. Technol. Adv. Eng., vol. 5, no. 1, pp. 567–571, 2015.

[12] R. A. Haraty, M. Dimishkieh y M. Masud, “An enhanced k-means clustering algorithm for pattern discovery in healthcare data”. Int. J. Distrib. Sens. Networks, vol. 2015, p. 11, 2015.

[13] Hartono, O. S. Sitompul, Tulus y E. B. Nababan, “Optimization Model of K-Means Clustering Using Artificial Neural Networks to Handle Class Imbalance Problem”. IOP Conf. Ser. Mater. Sci. Eng., vol. 288, no. 1, 2018.

[14] E. Sanez, “Minería de datos para una estrategia de medicina preventiva más robusta en una institución de salud pública del estado de Sonora”. Universidad de Sonora, 2018 .