Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 171-174 recibido 30-03-2019 revisado 25-04-2019 aceptado 14-05-2019

EXTRAE: EXTRacción de Asociaciones entre Enfermedades y otros conceptos médicos

EXTRAE: EXTRaction of Associations between Diseases and Other Medical Concepts

Lourdes Araujo[1], **Juan Martinez-Romo**[1], **Andrés Duque**[1], **Fernando López-Ostenero**[1], **Ricardo Sanchez de Madariaga**[2], **Adolfo Muñoz Carrero**[2], **Mario Pascual Carrasco**[2]
[1]Univ. Nacional de Educación a Distancia (UNED), 28040 Madrid
[2]Inst. de Salud Carlos III (ISCIII) 28029 Madrid
lurdes,juaner@lsi.uned.esaduque@scc.uned.es,fernando@lsi.uned.es
ricardo.sanchez@isciii.es adolfo.munoz@isciii.es mario.pascual@isciii.es

Resumen: El propósito de este proyecto es investigar en la mejora de las técnicas de extracción de Reglas de Asociación (RA) entre enfermedades, o entre enfermedades y otros conceptos médicos. Estas reglas permiten representar el conocimiento médico subyacente a un conjunto de Historias Clínica Electrónica (HCE). Concretamente nos planteamos explorar técnicas semisupervisadas que nos permitan alcanzar resultados equiparables a los de las técnicas supervisadas con una mínima supervisión. El proyecto se propone realizar avances significativos en la selección de reglas de asociación relevantes en el dominio de la salud, que pueden tener una alta aplicabilidad en la ayuda al diagnostico y en la prevención de enfermedades.
Palabras clave: dominio médico, extracción de información, reglas de asociación

Abstract: This project aims to improve the techniques for extracting Association Rules (AR) between diseases, or between diseases and other medical concepts. These rules allow the representation of medical knowledge underlying a set of Electronic Medical Records (EHR). Particularly, we plan to explore semi-supervised techniques that allow us to achieve similar results to those obtained using supervised techniques, but requiring minimal supervision. The project intends to make significant progress in the selection of relevant AR, which may be applied in the health domain for developing diagnostic help systems, or for disease prevention.
Keywords: medical domain, information extraction, association rules

1 Introducción

Los profesionales de la salud disponen en la actualidad de acceso a la Historia Clínica Electrónica (HCE) de los pacientes. La disponibilidad de información precisa, completa y estructurada permite mejorar considerablemente la toma de decisiones. Sin embargo, cada vez es más difícil tomar estas decisiones dado el gran volumen de datos que ha de considerarse. Este volumen dificulta encontrar manualmente relaciones que pueden ser utilizadas en la extracción de conocimiento. En este proyecto nos proponemos diseñar algoritmos que ayuden a la identificación de relaciones relevantes entre distintas enfermedades. Esta información es muy útil para realizar nuevos diagnósticos, probar nuevos tratamientos o fármacos, para prever la posible evolución de la enfermedad, etc. En la actualidad los médicos tienen que basarse en su experiencia para encontrar estas relaciones. El problema se hace prácticamente intratable cuando el especialista quiere abordar no sólo su área de especialización, sino también otras. Por esta razón sería muy útil disponer de un sistema que realice una preselección de las relaciones entre enfermedades y se las proponga a los especialistas en salud, para su consideración. Muchas enfermedades comparten uno, o varios aspectos, como síntomas, evolución, tratamiento, etc., pero esto no siempre significa que exista una relación entre ellas. Por ello, lo que propo-

ISSN 1135-5948. DOI 10.26342/2019-63-23

nemos es un sistema capaz de detectar relaciones entre enfermedades que se pueden considerar significativas. La significatividad vendrá dada por la coincidencia de aspectos más allá de la casualidad que se capturará definiendo un modelo estadístico apropiado. Las relaciones entre distintas enfermedades se pueden establecer en base a distintos patrones, separada o conjuntamente: aparición conjunta, síntomas comunes, similitudes de tratamientos, etc. Estas relaciones entre enfermedades se pueden codificar como Reglas de Asociación (RA), que se pueden considerar formas de representar el conocimiento médico subyacente en el conjunto de HCE almacenadas en el repositorio de información clínica. Sin embargo, la extracción de RA no resulta un proceso evidente o inmediato a partir del repositorio de HCE, por dos razones fundamentales. Primero, como se ha apuntado antes, la producción de relaciones entre enfermedades o RA por parte de los algoritmos de Aprendizaje Máquina (AM) no significa que estas relaciones sean significativas, pues pueden haber sido producidas por el ruido o la casualidad. Por otra parte, el conocimiento médico codificado en forma de RA debe ser ratificado por la experiencia empírica de los profesionales médicos. Este es un proceso extremadamente largo, complejo y costoso, si se tiene en cuenta que estos profesionales disponen de recursos de tiempo y esfuerzo limitados. Por lo tanto, es imprescindible la proposición a la comunidad científica médica de RA que tengan la máxima probabilidad a priori de ser validadas y añadidas al corpus de conocimiento médico establecido. En este proyecto estamos diseñando algoritmos semisupervisados (con la mínima intervención médica inicial posible, ya que este es un recurso muy costoso) para producir listas de RA plausibles ordenadas por la máxima probabilidad a priori de ser ciertas, de forma que a los profesionales médicos se les presenten únicamente las nuevas RA mejor colocadas en esta ordenación óptima, y así también se optimice el uso de un recurso tan preciado como el tiempo y el esfuerzo de estos profesionales.

2 Grupos Involucrados

Los grupos implicados en el proyecto son:

- Grupo NLP&IR[1] de la Universidad Nacional de Educación a Distancia (UNED). Dispone de una amplia experiencia en Acceso Inteligente a la Información y Adquisición y Representación de Conocimiento Léxico, Gramatical y semántico. Tiene una amplia trayectoria en la realización de proyectos de investigación.
- Grupo del Instituto de Salud Carlos III (ISCIII), con amplia experiencia en aprendizaje máquina de HCE estandarizada (norma UNE-EN ISO 13606), participación en proyectos de investigación y publicaciones científicas.

El grupo del Instituto de Salud Carlos III cuenta con datos de historiales clínicos (HCE) anonimizados con los que trabajar que suponen una excelente fuente de información para poder abordar el proyecto. Los extractos de HCE normalizados han sido transferidos por cinco hospitales españoles que han colaborado o colaboran en la actualidad con el ISCIII. Estos extractos han sido normalizados según el estándar UNE-EN ISO 13606 usando arquetipos diseñados por profesionales médicos de esos hospitales. Esta información médica está estructurada y sistematizada según el conocimiento médico de esos arquetipos. Esta fuerte estructuración y organización de la información médica la hace óptima para extraer de ella las reglas de asociación obtenidas por los algoritmos semisupervisados de AM. Este grupo cuenta así mismo con experiencia en la definición de reglas de asociación básicas.

Por su parte el grupo de la UNED cuenta con experiencia en métodos de aprendizaje automático que han aplicado en distintos problemas, y en particular en el dominio biomédico. Cuentan así mismo con experiencia en técnicas estadísticas que permitirán asignar probabilidades y establecer un ranking en las reglas de asociación que se quieren generar de forma automática.

3 Antecedentes

Uno de los modelos más estudiados en el ámbito de la minería de datos es el de las reglas de asociación (Agrawal, Imieliński, y Swami, 1993; Witten, Frank, y Hall, 2011). Este modelo se utiliza habitualmente para descubrir hechos que ocurren de forma común dentro de un mismo conjunto de datos. Por ejemplo, Witten, Frank, y Hall (2011) emplearon las reglas de asociación para descu-

[1] http://nlp.uned.es/

brir las relaciones entre los datos recopilados a gran escala en los sistemas de terminales de punto de venta (TPV) de unos supermercados. De esta forma, se extraían relaciones que indicaban por ejemplo, que si un usuario compraba cebollas y verduras al mismo tiempo, era muy probable que la carne también estuviera incluida en su cesta de la compra. Estas reglas de asociación representan las relaciones ocultas entre los datos de un conjunto. Estas son implicaciones del tipo $X \rightarrow Y$, donde X e Y son conjuntos de ítems disjuntos. Dentro de las reglas de asociación se manejan dos medidas que determinan la intensidad de la relación: el soporte y la confianza. El soporte de una regla se define como la frecuencia de dicha regla dado un conjunto de ítems, mientras que la confianza mide la frecuencia con la que los ítems del conjunto Y aparecen en las transacciones que contienen X. Estas dos medidas se utilizan con el fin de minimizar el hecho de encontrar patrones por casualidad. Para que la regla sea considerada fuerte, debe sobrepasar los umbrales establecidos de soporte y confianza, determinados por el experto para el conjunto de datos. Uno de los primeros trabajos (Stilou et al., 2001) que emplearon este tipo de reglas lo hicieron sobre una base de datos médica con información de pacientes diabéticos con el objetivo de identificar nuevos patrones hasta entonces no contemplados. Los autores consideraron los resultados como prometedores y sentaron las bases de la investigación de estos métodos en medicina.

Existen trabajos como el de Rashid, Hoque, y Sattar (2014) en el que tratan de encontrar co-ocurrencias entre las enfermedades padecidas por un paciente analizando un repositorio de informes clínicos. El sistema creado era capaz de predecir las co-relaciones entre enfermedades primarias (aquellas por las que el paciente visita al doctor) y secundarias, que son otras enfermedades asociadas a las primarias y padecidas por el mismo paciente. En la misma línea se encuentra el trabajo de Kang'ethe y Wagacha (2014) en el que sobre una base de 98.000 informes clínicos de pacientes de Estados Unidos y usando las reglas de asociación obtuvieron un nivel mínimo de confianza del 56-76 % y una co-ocurrencia del 90 % en la búsqueda de patrones de comorbilidad. El autor también uso la codificación ICD-9, algo que ayudó en la tarea de limitar el conjunto de posibles grupos de diagnóstico. La validación de reglas de asociación por profesionales médicos es un problema complejo y prolongado en el tiempo que requiere una cuidada selección de las potenciales reglas candidatas. Esto se debe al alto coste del tiempo y del esfuerzo de estos profesionales y de los estudios y ensayos clínicos. Por lo tanto es indispensable la optimización de los resultados producidos por los algoritmos de aprendizaje máquina aplicados a los repositorios de HCE normalizada. Para ello estamos explorando algoritmos de aprendizaje semi-supervisados, de forma que la intervención previa de los profesionales sanitarios sea minimizada. Estos algoritmos serán a su vez optimizados para que las listas de reglas de asociación que produzcan estén ordenadas según las probabilidades mayores de ser aceptadas por la comunidad médica internacional.

4 Objetivos

El principal objetivo del proyecto es la extracción de relaciones entre enfermedades que puedan ser inducidas a partir de los datos recogidos en informes médicos (HCE estandarizada). Estás relaciones deben estar ponderadas con probabilidades de forma que puedan indicar a los profesionales de la salud si existe un riesgo alto de observar una enfermedad futura en un paciente. Los objetivos específicos del proyecto son:

- Preparación de los datos de trabajo. Esto incluye la preparación de los textos, eliminando erratas, mayúsculas, etc. Se incluye también la generación de reglas de asociación utilizando algoritmos clásicos del área, como FP-growth (Han, Pei, y Yin, 2000), para generar las reglas candidatas. Parte de estas reglas candidatas han sido evaluadas manualmente por un doctor.
- Diseño e implementación de técnicas de aprendizaje automático semi-supervisadas para la extracción de relaciones entre enfermedades. Se incluye aquí la exploración de algoritmos no supervisados y supervisados que nos guíen en el diseño del algoritmo semisupervisado.
- Evaluación, comparativa y optimización de los resultados obtenidos con distintas técnicas.

- Obtención, evaluación y validación de resultados.

5 Situación actual

El proyecto, de dos años de duración, se encuentra ya en un estado avanzado de desarrollo. Contando con la colaboración de un doctor, se han anotado manualmente 1000 reglas de asociación, indicando si son ciertas o falsas. Estas reglas han permitido evaluar sistemas basadas en distintos enfoques. Se ha diseñado un sistema no supervisado basado en la significatividad estadística de las reglas. Se han construido también sistemas supervisados utilizando las reglas anotadas manualmente para entrenar. Finalmente se está diseñando un sistema semi-supervisado combinando propiedades del sistema no supervisado y del supervisado.

Agradecimientos

Este trabajo ha sido parcialmente financiado por los proyectos EXTRAE (IMIENS 2017) y PROSA-MED (TIN2016-77820-C3-2-R).

Bibliografía

Agrawal, R., T. Imieliński, y A. Swami. 1993. Mining association rules between sets of items in large databases. En *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, SIGMOD '93, páginas 207–216, New York, NY, USA. ACM.

Han, J., J. Pei, y Y. Yin. 2000. Mining frequent patterns without candidate generation. *SIGMOD Rec.*, 29(2):1–12, Mayo.

Kang'ethe, S. M. y P. W. Wagacha. 2014. Extracting diagnosis patterns in electronic medical records using association rule mining. *International Journal of Computer Applications*, 108(15):19–26, December. Full text available.

Rashid, M. A., M. T. Hoque, y A. Sattar. 2014. Association rules mining based clinical observations. *CoRR*, abs/1401.2571.

Stilou, S., P. D. Bamidis, N. Maglaveras, y C. Pappas. 2001. Mining association rules from clinical databases: An intelligent diagnostic process in healthcare. En *MEDINFO 2001 - Proceedings of the 10th World Congress on Medical Informatics, September 2-5, 2001, London, UK*, páginas 1399–1403.

Witten, I. H., E. Frank, y M. A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edición.