CLASIFICACIÓN AUTOMÁTICA DE ESTUDIOS EPIDEMIOLÓGICOS REFERENTES A DISTINTOS TIPOS DE CANCER UTILIZANDO TÉCNICAS DE MINERÍA DE TEXTO Y META-ESTIMADORES

MOUNIER, Mónica R.[1,2]; ACOSTA, Karina B.[1]; FAVRET, Fabián[2]; ZAMUDIO, Eduardo[1]

1 - Facultad de Ciencias Exactas Químicas y Naturales, Universidad Nacional de Misiones (UNaM), Félix de Azara 1552, Posadas, Misiones-Argentina. Tel: +54 (0376) - 4422186
2 - Universidad Gastón Dachary (UGD), Avda. López y Planes 6519, Posadas, Misiones-Argentina. Tel: +54 (0376) - 4438677

E-mail: monicamounier@fceqyn.unam.edu.ar

Eje Temático: "Ingeniería y Tecnología"

En la última década se ha visto un enorme crecimiento en la cantidad de datos biomédicos experimentales y computacionales, específicamente en las áreas de genómica y proteómica. Este crecimiento ha aumentado el número de publicaciones biomédicas referentes a hallazgos en estudios epidemiológicos de tipo caso-control, que reflejan la asociación de polimorfismos de nucleótidos simples (SNPs) y su asociación a distintos tipos de cáncer. Debido a ello, hay un gran interés por parte de la comunidad científica en herramientas de minería para ayudar a clasificar la abundante documentación disponible, a fin de encontrar datos relevantes para tareas de análisis específicas. Los SNPs son variaciones de la secuencia de ácido desoxirribonucleico (ADN) que se producen cuando se altera un solo nucleótido (A, T, C o G) en el genoma humano. La minería de texto (MT) procesa la información no estructurada y extrae índices numéricos desde el texto, a fin de que sea accesible para algoritmos de minería de datos.

El objetivo principal ha sido el desarrollo e implementación de una herramienta bioinformática de clasificación automática de estudios epidemiológicos de tipo caso-control referentes a SNPs relacionados a distintos tipos de cáncer utilizando técnicas de MT, a partir de sus metadatos.

Para el presente trabajo ha sido adaptada la metodología CRISP-DM, cuyas etapas son: recuperación y pre-procesamiento de metadatos, representación de datos y descubrimiento del conocimiento.

Fue elaborado un *dataset* a partir de los metadatos de 198 citas bibliográficas de artículos científicos elegidos aleatoriamente, y clasificadas por el experto en dos categorías: "Asociados" (169 artículos) y "No Asociados" (29 artículos). Un problema intrínseco es el desbalanceo de clases, dado que la mayoría de los estudios epidemiológicos reflejan asociaciones de los SNP a las enfermedades y no lo contrario.

La herramienta desarrollada consta de los siguientes módulos: consulta, recuperación, pre-procesamiento, clasificación, visualización y retroalimentación. Para su implementación fueron utilizadas las siguientes tecnologías: *Biopython*, *E-utilities* y *genenames.org Rest Web Service*, así también como *Django* para el desarrollo de la interfaz de consulta web. Para la representación de los metadatos de los artículos fue utilizado el *Term Frequency - Inverse Document Frequency* (TF-IDF) de los unigramas de los mismos. Para la clasificación fue utilizado el meta-estimador *Bagging*, para tres técnicas de clasificación: *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN) y *Naives Bayes* (NB), utilizando el 60 % del

dataset para entrenamiento y el 40 % restante para validación, donde cada meta-estimador fue entrenado y validado sobre el mismo subconjunto de datos para comparar los resultados obtenidos.

Los resultados obtenidos fueron superiores para el meta-estimador *Bagging* con NB, alcanzando una exactitud del 0.98 %, lo cual se obtuvo a partir de los resultados de la matriz de confusión obtenida a partir del subconjunto de validación conformado por 79 artículos, en donde de 65 artículos correspondientes a la categoría “Asociados”, 64 artículos fueron clasificados correctamente, y de los 14 artículos pertenecientes a la categoría “No Asociados”, 6 fueron clasificados correctamente. Así también, el mismo meta-estimador ha alcanzado una precisión de 0.88, una cobertura de 0.89, y un F1-Score de 0.87.

***Palabras clave**: Bioinformática, Minería de Textos, Meta-estimadores, Polimorfismos, Clasificación Automática, Estudios Epidemiológicos.*