POLIMORFISMOS DE NUCLEÓTIDOS SIMPLES RELACIONADOS AL RIESGO DE ENFERMEDADES: CLASIFICACIÓN AUTOMÁTICA DE ESTUDIOS EPIDEMIOLÓGICOS DE TIPO CASO-CONTROL UTILIZANDO TÉCNICAS DE MINERÍA DE TEXTO

Mónica R. Mounier[1,2,a], Karina B. Acosta[1,b], Fabián Favret[2,c], Eduardo Zamudio[1,d] , Diego A. Godoy[2,e], Juan de Dios Benítez[2,f]

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

[a]monicamounier@fceqyn.unam.edu.ar, [b]acostakb2505@gmail.com, [c]fabianfavret@citic.ugd.edu.ar, [d]eduardo.zamudio@ fceqyn.unam.edu.ar, [e]diegodoy@citic.ugd.edu.ar, [f]juan.benitez@citic.ugd.edu.ar

RESUMEN

En este trabajo de investigación se presenta una herramienta bioinformática que permite clasificar automáticamente artículos científicos referentes a estudios epidemiológicos de tipo caso-control concernientes a Polimorfismos de Nucleótidos Simples (SNPs), presentes en genes, y su asociación a distintos tipos de cáncer, y otras enfermedades genéticas de interés para el experto mediante la utilización de técnicas de minería de texto (MT), así también como la implementación del meta-estimador *Bagging* para tres técnicas de clasificación: *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), y *Naives Bayes* (NB). La clasificación se realiza a partir de los metadatos de dichos artículos, los cuales están disponibles en el *National Center for Biotechnology Information* (NCBI).

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

CONTEXTO

El trabajo presentado en este artículo tiene como contexto marco el proyecto de investigación denominado "Simulación en las Tics: Diseño de Simuladores de Procesos de Desarrollo de Software Ágiles y Redes De Sensores Inalámbricos para la Industria y la Academia", registrado actualmente en la Secretaría de Investigación y Desarrollo de la Universidad Gastón Dachary (UGD) con el número Código IP A07003 y radicado en el Centro de Investigación en Tecnologías de la Información y Comunicaciones de dicha universidad.

El mismo fue incorporado como proyecto aprobado en el llamado a presentación interna de la UGD de proyectos de investigación N°7 mediante la Resolución Rectoral 07/A/17 y es una continuidad del Proyecto Simulación como herramienta para la mejora de los procesos de software desarrollados con metodologías ágiles utilizando dinámica de sistemas, R.R. UGD N° 18/A/14 y R.R. UGD N° 24/A/15.

1. INTRODUCCIÓN

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 discusiones sobre hallazgos. 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 y útiles para tareas de análisis específicas [1].

Particularmente, los investigadores en el área biomédica, en adelante llamados expertos, presentan, como resultado de sus investigaciones y hallazgos, artículos científicos no estructurados. Dichos artículos son utilizados posteriormente por otros expertos para el estudio, diagnóstico, tratamiento y/o prevención de enfermedades. El inmenso cuerpo y rápido crecimiento del corpus biomédico han llevado a la aparición de un gran número de técnicas de MT destinadas a la extracción automática de conocimiento [2].

Actualmente, existe la necesidad de disponer de una herramienta que permita clasificar automáticamente artículos científicos referentes a estudios epidemiológicos de tipo caso-control, que reflejen la asociación de SNPs, presentes en genes, y su asociación a enfermedades específicas. La fuente principal de metadatos de los artículos a clasificar se encuentra disponible en el NCBI [3]. Hoy en día, dicha clasificación es realizada manualmente por el experto, lo cual resulta notablemente ineficiente, dada la cantidad de tiempo necesaria para llevarla a cabo, sumado a las constantes actualizaciones de la bibliografía disponible.

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. Los SNPs representan alrededor del 90% de toda la variación genética humana, que se producen cada 100-300 bases a lo largo del genoma humano (3 mil millones de bases), aunque su densidad varía entre las regiones [4].

La bioinformática consiste en la investigación, desarrollo y/o aplicación de herramientas computacionales y enfoques para ampliar el uso de datos biológicos, médicos, de comportamiento o de salud, incluidas las de adquirir, almacenar, organizar, archivar, analizar y visualizar esos datos [5].

La MT procesa la información no estructurada y extrae índices numéricos desde el texto, a partir de lo cual hace a la información accesible para varios algoritmos de minería de datos. Básicamente, la MT convierte el texto en números los cuales pueden luego ser incluidos en otros análisis [6].

Las tecnologías de MT básicas pueden ser utilizadas individualmente o en forma conjunta, dependiendo del problema de minería a resolver. Dichas tecnologías son las siguientes [7]:

- Recuperación de información;
- Extracción de información;
- Categorización o clasificación;
- *Clustering*;
- Generación Automática de Resúmenes.

En la recuperación de información, los documentos son generalmente identificados por un conjunto de términos o palabras clave que son usadas colectivamente para representar su contenido [8]. Existen muchas técnicas de representación de textos, siendo algunas de las más utilizadas las siguientes [9, 10]:

- *Inverse Document Frequency* (IDF);
- *Term Frequency – Inverse Document Frequency* (TF-IDF).

El método *Bagging* es un meta-estimador en el cual los clasificadores individuales son entrenados en paralelo a partir de muestras con reemplazo obtenidas aleatoriamente del mismo conjunto de entrenamiento. Para construir un comité de n clasificadores (siendo n el número de clasificadores a utilizar), se debe elegir la forma de combinar los resultados de dichos clasificadores, siendo el método más simple el voto mayoritario en el cual la categoría asignada a un documento es aquella que fuera elegida por la mayoría de los clasificadores, donde n debe ser un número impar [11]. En la Figura 1 se muestra el esquema de funcionamiento del meta-clasificador *Bagging*.

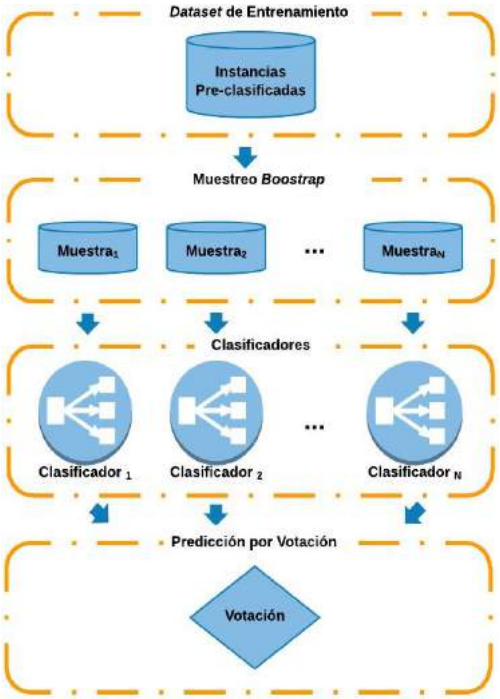

Figura 1. Esquema del Meta-estimador Bagging

2. LÍNEAS DE INVESTIGACIÓN

En esta investigación el objetivo es 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, y otras enfermedades genéticas de interés para el experto utilizando técnicas de MT, a partir de sus metadatos.

Objetivos Específicos

- Identificar los metadatos de recuperación y clasificación necesarios para la elaboración de la herramienta bioinformática;
- Seleccionar las técnicas adecuadas para el tratamiento de los datos recuperados;
- Obtener una base de instancias pre-clasificadas por el experto a partir de la información relevante recuperada a través del NCBI relacionada a estudios epidemiológicos de tipo caso-control referentes SNPs
- Diseñar e implementar la herramienta bioinformática para la clasificación automática de estudios epidemiológicos de tipo caso-control referentes SNPs relacionados a enfermedades de interés para el experto

3. RESULTADOS

Para la elaboración del *dataset* fueron consideradas aleatoriamente los metadatos de 198 citas bibliográficas de artículos científicos de interés para el experto, clasificadas por el mismo en dos categorías: “Asociados” (169 artículos) y “No Asociados” (29 artículos). Siendo un problema intrínseco el desbalanceo de las

clases, dado que la mayoría de los estudios reflejan asociaciones de los SNP a las enfermedades y no lo contrario.

Para el presente trabajo ha sido realizada la adaptación de una metodología CRISP_DM [12], así también como la metodología propuesta en un trabajo de investigación aplicado a MT [13]. Las etapas que componen la metodología son:

1. Recuperación de metadatos;
2. Pre-procesamiento de metadatos;
3. Representación de los datos;
4. Descubrimiento de conocimiento.

En la Figura 2 se presenta la herramienta que ha sido desarrollada en el presente trabajo. La misma está conformada por los siguientes módulos:

1. Módulo de consulta;
2. Módulo de recuperación;
3. Módulo de pre-procesamiento;
4. Módulo de clasificación;
5. Módulo de visualización;
6. Módulo de retroalimentación.

Las herramientas bioinformáticas utilizadas fueron:

• Biopython: Herramientas de libre acceso para la biología computacional y bioinformática en Python.

• E-utilities: Conjunto de programas que proporcionan una interfaz de consulta de Entrez del NCBI.

• genenames.org Rest Web Service: Servicio web proporcionado por el *HUGO Gene Nomenclature Commitee*.

Para la representación de los metadatos de los artículos fue utilizado TF-IDF de los unigramas de los mismos.

Para la clasificación fue utilizado el meta-estimador *Bagging*, por ser el más adecuado para *dataset* desbalanceados [14], 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.

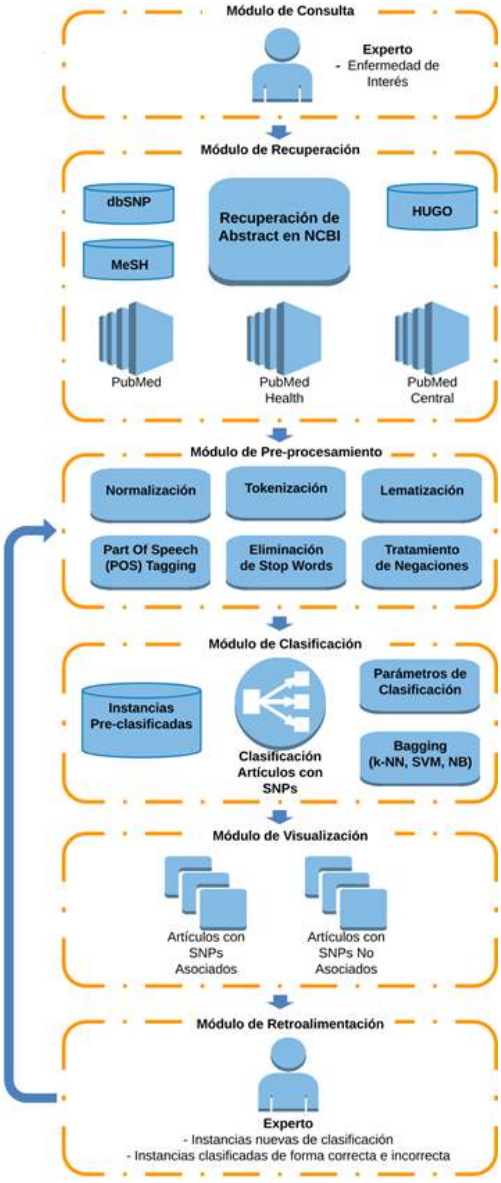

Figura 2. Esquema del Meta-estimador

En la Tabla 1 se presentan los resultados obtenidos de la clasificación para los tres meta-clasificadores utilizando las tres técnicas de clasificación planteadas, las cuales fueron entrenadas y validadas sobre el mismo subconjunto de datos, a fin de comparar el comportamiento de cada meta-clasificador.

Tabla 1. Resultados obtenidos

	***Bagging* SVM**		***Bagging* KNN**		***Bagging* NB**	
Matriz de Confusión						
	A	NA	A	NA	A	NA
A	65	0	64	1	64	1
NA	10	4	13	1	8	6
Medidas de Evaluación						
Precisión	0.89		0.77		0.88	
Cobertura	0.87		0.82		0.89	
F1-Score	0.84		0.76		0.87	
Exactitud	0.97		0.94		0.98	

Donde A: Asociados; NA: No Asociados

Como puede verse en Tabla 1, los resultados obtenidos fueron superiores para el *Bagging* con NB, alcanzando una exactitud del 0.98 %.

4. FORMACIÓN DE RECURSOS HUMANOS

El trabajo corresponde a un trabajo de fin de carrera de grado de Ingeniería en Informática. El equipo de investigación está formado por un Doctor en Ciencias de la Computación, una Doctora en Ciencias Biológicas, ambos, docentes-investigadores de la UNaM, un Doctor en Tecnologías de la Información (TI) y Comunicación, un maestrando en TI, un maestrando en Redes de datos y ocho estudiantes en período de realización de trabajos finales de grado en el contexto de las carreras de Licenciatura en Sistemas de Información y de Ingeniería en Informática de la UGD. Actualmente, el número de tesinas de grado aprobadas en el contexto de este proyecto, es de cinco, y otras tres en proceso de desarrollo.

5. REFERENCIAS

[1] H. Shatkay and R. Feldman, "Mining the biomedical literature in the genomic era: an overview", Journal of computational biology, vol. 10, no. 6, pp. 821-855, 2003.

[2] F. Zhu et al., "Biomedical text mining and its applications in cancer research", Journal of biomedical informatics, vol. 46, no. 2, pp. 200-211, 2013.

[3] (2017, Feb.) National Center for Biotechnology Information (NCBI). [Online]. http://www.ncbi.nlm.nih.gov

[4] J. E. Lee, J. H. Choi, J. H. Lee, and M. G. Lee, "Gene SNPs and mutations in clinical genetic testing: haplotype-based testing and analysis", Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, vol. 573, no. 1, pp. 195-204, 2005.

[4] M. Huerta, G. Downing, F. Haseltine, B. Seto, and Y. Liu, "NIH working definition of bioinformatics and computational biology", US National Institute of Health, 2000.

[6] R. Agrawal and M. Batra, "A detailed study on text mining techniques", International Journal of Soft Computing and Engineering, vol. 2, no. 6, pp. 118-121, 2013.

[7] S. Dang and P. H. Ahmad, "Text Mining: Techniques and its Application", International Journal of Engineering & Technology Innovations, pp. 2348-0866, 2014.

[8] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of TF* IDF, LSI and multi-words for text classification", Expert Systems with Applications, vol. 38, no. 3, pp. 2758-2765, 2011.

[9] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval", Journal of documentation, vol. 28, no. 1, pp. 11–21, 1972.

[10] K. S. Jones, "IDF term weighting and IR research lessons" Journal of documentation, vol. 60, no. 6, pp. 521-523, 2004.

[11] L. Breiman, "Bagging predictors" Machine learning, vol. 24, no. 2, pp. 123-140, 1996.

[12] P. Chapman et al., "CRISP-DM 1.0 Step-by-step data mining guide", 2000.

[13] C. Gálvez, "Minería de textos: la nueva generación de análisis de literatura científica en biología molecular y genómica" Encontros Bibli: rev. eletrônica de bib. e ciência da inf., vol. 13, no. 25, pp. 1518-2924, 2008.

[14] G. Collell, D. Prelec, K. Patil, “A simple plug-in bagging ensemble based on threshold-moving for classifying binary and multiclass imbalanced data”, Neurocomputing, vol. 275, 330-340, 2018.