**SIMPOSIO “TRANSFERENCIA DE CONOCIMIENTO EN TECNOLOGÍAS DE LA INFORMACIÓN”**

**Modelos de multiclasificación aplicados al análisis del comportamiento de enfermedades prioritarias según datos del Ministerio de Salud Pública del Ecuador**

***Multiclassification models applied to the analysis of the behavior of priority diseases according to data from the Ministry of Public Health of Ecuador***

**Oscar José Alejo Machado1, Rafael Esteban Bello Pérez2, María Matilde García Lorenzo3 y Maikel Yelandi Leyva Vázquez4**

1- Oscar José Alejo Machado. Instituto Superior Universitario Bolivariano de Tecnología, Ecuador. E-mail: oalejo@bolivariano.edu.ec

2- Rafael Esteban Bello Pérez. Universidad Central “Marta Abreu” de Las Villas, Cuba. E-mail: rbellop@uclv.edu.cu

3- María Matilde García Lorenzo. Universidad Central “Marta Abreu” de Las Villas, Cuba. E-mail: mmgarcia@uclv.edu.cu

4- Maikel Yelandi Leyva Vázquez. Universidad Regional Autónoma de los Andes, Ecuador. E-mail: ub.c.investigacion@uniandes.edu.ec

**Resumen:** En el presente trabajo se estudian las características de prevalencia de algunas enfermedades llamadas prioritarias. Este tipo de enfermedades, aunque tienen causas diferentes, donde algunas son infecciosas y otras son crónicas no transmisibles, tienen en común que constituyen un problema importante para la salud pública de las sociedades modernas y a la vez su alivio o adquisición dependen de la conducta de los pacientes. El Ministerio de Salud Pública de Ecuador tiene archivados en sus bases de datos un volumen importante de historias clínicas de los pacientes con tales enfermedades. Estos repositorios y registros contienen informaciones tácitas que de ser extraídas pueden resultar en elementos claves para trazar políticas de salud más eficientes y eficaces, tanto desde el punto de vista humano, como desde el punto de vista económico y social. Estos datos no son unidimensionales, contienen elementos en forma de texto, imágenes, entre otros. Para extraer la información se utilizan métodos de multiclasificación o de multietiqueta, donde a cada ejemplo o caso se le asignan múltiples etiquetas. El objetivo de este artículo es la obtención de esta información valiosa, con ayuda de métodos de multietiquetado. Se utiliza el software MEKA para realizar los estudios.

***Abstract:*** *In the present work, the prevalence characteristics of some so-called priority diseases are studied. These types of diseases, although they have different causes, where some are infectious and others are chronic non-communicable, have in common that they constitute an important problem for the public health of modern societies and, at the same time, their alleviation or acquisition depends on the behavior of the patients. patients. The Ministry of Public Health of Ecuador has an important volume of medical records of patients with such diseases on file in its databases. These repositories and registries contain tacit information that, if extracted, can result in key elements to draw up more efficient and effective health policies, both from the human point of view, as well as from the economic and social point of view. This data is not one-dimensional, it contains elements in the form of text, images, among others. To extract the information, multiclassification or multilabel methods are used, where each example or case is assigned multiple labels. The goal of this article is to obtain this valuable information, with the help of multilabeling methods. MEKA software is used to carry out the studies*.

**Palabras Clave:** Multiclasificación; Aprendizaje multietiqueta; Minería de datos; MEKA; Enfermedades prioritarias; Salud Pública.

***Keywords:*** *Multiclassification; Multilabel learning; Data mining; MEKA; Priority diseases; Public health*.

**1. Introducción**

En las sociedades modernas existe un grupo de enfermedades que su prevalencia en la población constituye un problema de salud que debe ser atendido por el estado. Algunas son transmisibles de origen viral o bacteriano, otras son crónicas no transmisibles. La adquisición de estas enfermedades y la mejoría de los pacientes que las padecen tienen una gran dependencia del comportamiento de ellos mismos.

El estudio de la prevalencia de tales enfermedades en Ecuador puede ser decisivo para trazar políticas más eficaces y eficientes para combatir estas enfermedades. Para ello, el Ministerio de Salud Pública nacional cuenta con bases de datos que contienen la historia clínica de estos pacientes, con datos que se expresan en forma de texto, imágenes, entre otros formatos, conteniendo una información valiosa del estado y evolución de estas personas en cuanto a su enfermedad, así como sus hábitos, factores de riesgo conductuales o no, entre otros.

El problema se convierte en la necesidad de extracción de la información y por consiguiente el conocimiento que permanece implícitamente dentro de tales bases de datos, y para ello el aprendizaje automatizado constituye un área fundamental de la computación para resolver dicha necesidad.

En el aprendizaje automatizado, la clasificación multietiqueta constituye una de las variantes de los problemas de clasificación. Para este tipo de clasificación a cada ejemplo o caso se le asignan múltiples etiquetas. [1]

La clasificación multietiqueta es una generalización de la clasificación multiclases, que consiste en el problema de categorizar un ejemplo con una etiqueta existiendo más de dos clases posibles: en el problema multietiqueta no existen restricciones con respecto a cuantas etiquetas se le asignan a un mismo caso.

La clasificación multietiqueta apareció con el problema de la categorización de texto donde cada documento podía pertenecer a diferentes tópicos simultáneamente y se ha generalizado a diferentes áreas, lo cual incluye el procesamiento de datos estructurados, semiestructurados o no estructurados. [2]

Formalmente, la clasificación multietiqueta es el problema de encontrar un modelo que permita reconocer dado un conjunto de vectores o entradas representadas por X, un vector binario Y (asignando un valor 0 o 1 para cada elemento o etiqueta de Y).

Existen múltiples métodos para resolver problemas de asignación de multietiquetas. En Tsoumakas y Katakis se clasifican en dos categorías fundamentales: métodos basados en algoritmos de adaptación y métodos de transformación de problemas [3]. La primera de estas categorías generaliza métodos de adaptación para tratar directamente el problema de multietiquetado, donde se adaptan métodos muy conocidos como son los árboles de decisión, reglas de clasificación o redes neuronales artificiales. Mientras que los métodos de transformación de problemas transforman el problema de multietiquetado en uno o varios problemas de etiquetado único, para después resolver estos por los métodos que se utilizan para este tipo de problemas más simples.

Entre los métodos utilizados para resolver estos problemas se pueden enumerar: Binary Relevance (BR), Classifier Chains (CC), el basado en Back-Propagation Neural Network (BPNN), Multilabel k Neighbourn Nearest (MLkNN), el basado en Back-Propagation Multilabel (BPMLL) y el basado en el método RAkEL, que constituye una mejora sustancial al método conocido como Label Powerset (LP), especialmente en dominios con muchas etiquetas. [3-12]

El presente artículo tiene como objetivo la obtención de información valiosa sobre las enfermedades priorizadas, con ayuda de métodos de multietiquetado. Se utiliza el software MEKA para realizar los cálculos. Este estudio tiene gran importancia para mejorar la calidad de la atención a los pacientes en el sistema de salud pública de Ecuador, en especial para mejorar el tratamiento de enfermedades priorizadas, de manera que estos mejoren en cuanto a eficacia médica y eficiencia de recursos materiales.

**2. Conceptos preliminares**

El aprendizaje multietiqueta se relaciona con el aprendizaje con ejemplos, donde cada uno de ellos se asocia a múltiples etiquetas, las que pueden pertenecer a un conjunto predefinido de etiquetas. El objetivo de la clasificación multietiqueta es la construcción de un modelo predictivo que provea una lista de etiquetas relevantes para la clasificación de ejemplos nuevos dados previamente. [5]

Este problema consistirá formalmente en los siguientes elementos:

Un espacio de ejemplos χ que consiste en tuplas de valores de tipos de datos (booleanos, discretos o continuos).

Formalmente, xi∈χ, xi = (xi1, xi2, …, xiD) donde D es el tamaño de la tupla, o el número de atributos descritos.

Un espacio de etiquetas $L$ = {λ1, λ2, …, λQ} que consta de en un conjunto de Q variables discretas.

Un conjunto de ejemplos E, donde cada ejemplo está formado por dos tuplas, una para el ejemplo y otra para la etiqueta. Es decir, $E= \left\{\left(x\_{i},y\_{i}\right)|x\_{i}\in χ, y\_{i}\in L, 1\leq i\leq N\right\}$ y N es el número de ejemplos.

Un criterio de calidad q que premia a los modelos con alta exactitud predictiva y baja complejidad.

El objetivo es encontrar una función h: χ→$2^{L}$ tal que h maximice q. Donde $2^{L}$ denota el conjunto potencia del conjunto de etiquetas.

En este artículo se aplican algoritmos contenidos en MEKA. MEKA es un framework de código abierto programado en Java que cuenta con interfaces que facilitan las aplicaciones prácticas con métricas de evaluación y herramientas para realizar experimentos y el desarrollo de clasificaciones multietiquetas.[1,13]

El método de relevancia binaria, o en inglés, Binary Relevance (BR) consiste en aplicar una estrategia de uno contra todos para convertir el problema de multietiqueta en varios problemas de clasificación binaria. Este método se relaciona estrechamente con el método de Cadena Clasificadora o Classifier Chain (CC), donde Q clasificadores binarios se vinculan a través de una cadena.

El espacio de características de cada vínculo en la cadena se extiende a una asociación 0/1 de todos los vínculos previos. La predicción de las etiquetas relevantes en el método CC es el mismo que en el método BR.

El método Label Power-set (LP) combina conjuntos completos de etiquetas en etiquetas únicas para formar problemas de una única etiqueta. De esa manera el problema se simplifica a uno de etiqueta única. Este algoritmo tiene el inconveniente que el espacio de subconjuntos de etiquetas posibles puede ser muy grande, es por ello que es necesario aplicar un método de podado que tiene en cuenta solamente aquellas etiquetas que ocurren más veces con respecto a un valor umbral.

El resultado del método Multi-Label k-nearest-neighbors es el mismo que en el algoritmo k-NN tradicional, la diferencia consiste en la determinación del conjunto de etiquetas de prueba.

El método RAndom k labEL sets (RAkEL) parte de pequeños conjuntos aleatorios de etiquetas para crear cada base del clasificador. Aprende de un clasificador de etiqueta única para la predicción de cada elemento en el conjunto potencia. La complejidad computacional del algoritmo se reduce aplicando métodos de podado. Este método ha demostrado ser eficiente en comparación con otros.

El uso del popular algoritmo de propagación hacia atrás (back-propagation) que se utiliza en las Redes Neuronales Artificiales, adaptado a problemas de multietiquetado, es la base de los métodos de back-propagation neural networks (BPNN) y back-propagation neural networks Multi-Label Learning (BPMLL).

Básicamente estos algoritmos reemplazan las funciones de error con una nueva función para capturar las características de aprendizaje multietiqueta. Las etiquetas que coinciden con el ejemplo se les dan una mayor evaluación que aquellas que no coinciden.

Otros se basan en métodos conocidos como el Multi-Label C4.5 que adapta el algoritmo C4.5 en el marco del multietiquetado. En este caso se utiliza la entropía.

**3. Resultados y discusión**

En la investigación se trabaja con nueve etiquetas que representan las enfermedades prioritarias sugeridas por las autoridades de salud. Estas etiquetas son:

1. TUBERCULOSIS
2. DENGUE-CHIKUNGUNYA-SIKA
3. VIH
4. DIABETES
5. CANCER
6. HIPERTENSIÓN
7. HEPATITIS A,B,C
8. INFLUENZA-NEUMONÍA
9. INFARTO AGUDO DEL MIOCARDIO

Considerando para su procesamiento la información basada en el código CIE10 [14].

Partiendo de los resultados alcanzados con la base de casos de 2015 para enfermedades prioritarias se trabajó con los casos reducidos a 20 rasgos y los nueves etiquetas correspondientes a las enfermedades prioritarias. No se contemplaron los rasgos referidos a procedimientos.

De forma similar al 2015 se trabaja con los métodos de clasificación multietiqueta: Binary Relevance(BR), Classifier Chains(CC) y RAKEL y como clasificador base se usa J48, los cuales son los mejores reportados en estudios anteriores.

Se analiza la construcción del árbol de decisión para distintos niveles de confianza considerándose 0,25, 0,5 y 0,7, como se resume en la Tabla 1.

**Tabla 1.** Comparación en cuanto a la función de *Hamming Loss* [15] entre métodos de multietiquetado con árboles de decisión con niveles de confianza 0,25, 0,5 y 0,7.

|  |  |  |  |
| --- | --- | --- | --- |
| Método / *Hamming Loss* | FC\_0.25 | FC\_0.5 | FC\_0.7 |
| BR | 0,081 | 0,077 | 0,079 |
| CC | 0,082 | 0,081 | 0,083 |
| RAKEL | 0,078 | 0,079 | 0,08 |

Fuente: Elaboración propia.

Por otro lado, partiendo de los resultados alcanzados con la base de casos del 2015 para enfermedades prioritarias se trabajará con los casos reducidos a 23 rasgos y las nueve etiquetas correspondientes a las enfermedades prioritarias. No se contemplaron los rasgos referidos a procedimientos y cantidad de procedimientos, los cuales suman seis (6), además de eliminar los rasgos nro\_id-unidad\_operativa, nacionalidad, provincia, cantón y parroquia de residencia del paciente, tipo de parroquia y tipología que son ocho (8) para un total de 14 rasgos eliminados, véase Tabla 2.

De forma similar al 2015 se trabaja con los métodos de clasificación multietiqueta: Binary Relevance(BR), Classifier Chains(CC) y RAKEL y como clasificador base se usa J48, los cuales son los mejores reportados en el informe anterior.

Dado que esta base contempla cuatro rasgos que no aparecieron en las bases de los años 2015 y 2016 se realizarán dos análisis: uno con 23 rasgos y otro con la base reducida a 19 rasgos y para esta última se eliminarán los rasgos: referencia, interconsulta, orientación sexual e identidad de género.

Se analiza la construcción del árbol de decisión para distintos niveles de confianza considerándose una base de casos con 23 rasgos.

**Tabla 2.** Comparación en cuanto a la función de *Hamming Loss* entre métodos de multietiquetado con árboles de decisión con niveles de confianza 0,25, 0,5 y 0,7.

|  |  |  |  |
| --- | --- | --- | --- |
| Método/*Hamming Loss* | FC\_0.25 | FC\_0.5 | FC\_0.7 |
| BR | 0,078 | 0,076 | 0,078 |
| CC | 0,079 | 0,079 | 0,08 |
| RAKEL | 0,077 | 0,078 | 0,079 |

Fuente: Elaboración propia.

No existen diferencias entre los valores de *Hamming Loss* que se alcanzan por parte de los distintos métodos, ni existen cambios significativos cuando se cambian el nivel de confianza (FC) del árbol que se debe obtener. Los resultados son muy ligeramente superiores para FC igual a 0,5.

En la Tabla 3, se muestra a continuación la comparación de métodos considerando 19 rasgos en la base de casos.

**Tabla 3.** Comparación en cuanto a la función de *Hamming Loss* entre métodos de de multietiquetado con árboles de decisión con niveles de confianza 0,25, 0,5 y 0,7.

|  |  |  |  |
| --- | --- | --- | --- |
| Método/Hamming Loss | FC\_0.25 | FC\_0.5 | FC\_0.7 |
| BR | 0,083 | 0,08 | 0,082 |
| CC | 0,084 | 0,083 | 0,084 |
| RAKEL | 0,081 | 0,081 | 0,083 |

Fuente: Elaboración propia.

No existen diferencias entre los valores de *Hamming Loss* que se alcanzan por parte de los distintos métodos, ni existen cambios significativos cuando se cambian el nivel de confianza (FC) del árbol que se debe obtener. Los resultados son ligeramente superiores para FC igual a 0,5.

**4. Conclusiones**

Se realizó el estudio de enfermedades que constituyen serios problemas de salud en las sociedades contemporáneas, en especial en Ecuador, se logra la adquisición de conocimiento vital para el sistema de salud de la provincia Guayas sobre la base de las enfermedades prioritarias estudiadas y la aplicación de métodos de clasificación multietiqueta.

De la experimentación realizada se concluyen que los métodos de mejores comportamiento son: Binary Relevance (BR), Classifier Chains (CC) y RAKEL, combinados con el clasificador base J48, tomando con base la evaluación con Hamming Loss

La selección de rasgos realizada sobre la base de casos de estudio no aporta diferencias significativas con respecto al tratamiento inicial una vez preprocesados los datos..

Finalmente se obtuvieron las relaciones entre las diferentes variables que permiten asociar algunas características de los pacientes como hábitos, situación social, entre otras, con la enfermedad que sufren. Por ejemplo, se descubrió que existe una relación entre el lugar de atención médica y la prevalencia de tuberculosis.

**5. Referencias bibliográficas**

1. Modi H, Panchal M. Experimental Comparison of Different ProblemTransformation Methods for Multi-Label Classificationusing MEKA. International Journal of Computer Applications 2012;59(15):10-5.

2. Zhang ML, Zhou ZH. Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization. IEEE Transactions on Knowledge and Data Engineering. 2006;18(10):1338–51.

3. Tsoumakas G, Katakis I. Multi label classiﬁcation: an overview. International Journal of Data Warehouse and Mining. 2007;3(3):1–13.

4. Elissef A, Weston J. A Kernel Method for Multi-Labelled Classification. En: Dietterich TG, Becker S, Ghahramani Z, editores. Advances in Neural Information Processing Systems; 2002. p. 681–7.

5. Madjarov G, Kocev D, Gjorgjevikj D, Dzeroski S. An extensive experimental comparison of methods for multi-label learning. Pattern Recognition. 2012;45(9):3084-104.

6. Pestian JP, Brew C, Matykiewicz P, Hovermale DJ, Johnson N, Cohen KB, et al. A shared task involving multi-label classification of clinical free text. Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing; 2007. Association for Computational Linguistics; 2007. p. 97-104.

7. Read J, Pfahringer B, Holmes G, Frank E. Classifier Chains for Multi-label Classification. Machine Learning Journal. 2011;85(3):333-59

8. Read J, Pfahringer B, Holmes G, editors. Multi-label classification using ensembles of pruned sets. 8th IEEE international conference on data mining; 2008. IEEE.

9. Read J, Perez-Cruz F. Deep learning for multi-label classification. arXiv preprint arXiv:150205988. 2014.

10. Tsoumakas G, Katakis I, Vlahavas I. Random k-Label set for Multilabel Classification. IEEE Transactions of Pattern Analysis and Machine Learning. 2007;23(7):1079-89.

11. Zaragoza JC, Sucar E, Morales E, Bielza C, Larranaga P. Bayesian chain classifiers for multidimensional classification. Twenty-second international joint conference on artificial intelligence; 2011. 2011.

12. Zhang ML, Zhou ZH. ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition. 2007;40(7):2038–48.

13. Read J, Reutemann P, Pfahringer B, Holmes G. Meka: A Multi-label/Multi-target Extension to Weka. Journal of Machine Learning Research 2016;17(1):667-71.

14. OMS. Disponible en: http://www.cie10.org/. Consultado el 10-1-2019. CIE-10 en español, descarga o consulta [updated CIE-10 en español, descarga o consulta; cited]; Available from.

15. Norouzi M, Fleet DJ, Salakhutdinov RR. Hamming distance metric learning. Advances in neural information processing systems; 2012. p. 1061-9.