ANALES DE LA
Academia de Ciencias de Cuba

CIENCIAS TÉCNICAS

Artículo original de investigación

Desarrollo de técnicas para el preprocesamiento y la predicción de problemas de clasificación multietiqueta

Marilyn Bello García [1*] https://orcid.org/0000-0002-8804-749X
Rafael E. Bello Pérez [1] https://orcid.org/0000-0001-5567-2638
Gonzalo Nápoles [2] https://orcid.org/0000-0003-1936-3701
Koen Vanhoof [3] https://orcid.org/0000-0001-7084-4223
María M. García Lorenzo [1] https://orcid.org/0000-0002-1663-5794
Yaumara Aguilera Calzadilla [4] https://orcid.org/ 0000-0003-3440-49542

[1] *Centro de Investigaciones de la Informática, Universidad Central Marta Abreu de Las Villas. Santa Clara, Cuba*
[2] *Departmento de Ciencia Cognitiva e Inteligencia Artificial, Universidad de Tilburg. Tilburg, Países Bajos*
[3] *Facultad de Negocios y Economía, Universidad de Hasselt. Hasselt, Bélgica*
[4] *Hospital Comandante Manuel Fajardo Rivero. Santa Clara, Cuba*

*****Autor para la correspondencia:* mbgarcia@uclv.cu

Revisores

Orestes Llanes Santiago
Universidad Tecnológica de La Habana.
La Habana, Cuba

Edel García Reyes
Geocuba Investigación y Consultoría.
La Habana, Cuba

Editor

Lisset González Navarro
Academia de Ciencias de Cuba.
La Habana, Cuba

Traductor

Darwin A. Arduengo García
Academia de Ciencias de Cuba.
La Habana, Cuba

RESUMEN

Introducción: La clasificación multietiqueta es una variante de la clasificación tradicional de etiqueta única, en la que un objeto ya no se clasifica exclusivamente por una etiqueta. En su lugar, este aprendizaje pretende asignar a un objeto una o más clases de etiquetas de un conjunto predefinido de clases. Dado que el aprendizaje multietiqueta se encuentra todavía en una fase temprana de desarrollo, en comparación con otras técnicas de clasificación, algunas técnicas actualmente disponibles para otros tipos de aprendizaje no se han desarrollado para este caso específico. **Métodos:** Tras un estudio de la literatura existente, los siguientes son algunos de los retos de investigación dentro de esta temática: medidas de calidad de los datos, métodos de reducción sobre conjuntos de datos multietiqueta, métodos de detección de valores atípicos, capas de agrupación para datos multietiqueta sin una organización topológica, métodos para tratar problemas de clasificación multietiqueta con características dispersas y técnicas de inteligencia artificial explicable para clasificadores neuronales multietiqueta. **Resultados:** Se proponen: a) Medidas de calidad de los datos multietiqueta (3); b) Métodos para reducir conjuntos de datos multietiqueta (6); c) Método que mide el grado de anomalía de un objeto en conjuntos de datos multietiqueta (1); d) Arquitectura neuronal profunda que utiliza capas de agrupación basadas en la asociación bidireccional (1); e) Sistema neuronal para resolver problemas de clasificación multietiqueta descritos por datos tabulares que pueden implicar características dispersas (1) y f) Adaptación al escenario multietiqueta de una técnica clásica de interpretabilidad post-hoc en redes neuronales (1). Conclusiones, los métodos propuestos le proporcionan a la comunidad científica novedosas técnicas de clasificación multietiqueta, haciendo posible un proceso de descubrimiento de conocimiento más eficiente y eficaz sobre datos multietiqueta.

Premio Nacional Anual de la Academia de Ciencias de Cuba, 2021

Palabras clave: clasificación multietiqueta; caracterización de los datos; preprocesamiento de los datos; proceso de aprendizaje; inteligencia artificial explicable

Development of techniques for pre-processing and prediction of multi-label classification problems

ABSTRACT

Introduction: Multi-label classification is a variant of traditional single-label classification, where an object is no longer classified by exclusively one label. Instead, this learning aims to assign one or more classes from a predefined set of classes to an object. Since multi-label learning is still in an early development stage compared to other classification techniques, some techniques currently available for other learning types have not been developed for this specific learning case. **Methods:** After a survey of the existing literature, the following are some research challenges within this topic: data quality measures, reduction methods on multi-label datasets, outlier detection methods, pooling layers for multi-label data without a topological organization, methods to deal with multi-label classification problems with sparse features, and Explainable Artificial Intelligence techniques for multi-label neural classifiers. **Results:** We propose: a) three measures of multi-label data quality, b) six methods for reducing multi-label datasets, c) a method that measures an object's anomaly degree in a multi-label dataset, d) a deep neural architecture using bidirectional association-based pooling layers, e) a neural system to solve multi-label classification problems described by tabular data that might involve sparse features, and f) an adaptation to the multi-label scenario of a classical post-hoc interpretability technique on neural networks. Conclusions, the proposed methods provide the scientific community with novel multi-label classification techniques, making possible a more efficient and effective knowledge discovery process on multi-label data.

Keywords: multi-label classification; data characterization; data pre-processing; learning process; explainable artificial intelligence

INTRODUCCIÓN

El desarrollo de un modelo de clasificación suele estar orientado a descubrir la correspondencia subyacente entre un conjunto de objetos de entrada caracterizados por un espacio de características multidimensional y la etiqueta de clase que se les asigna. Sin embargo, ¿qué ocurre si cada objeto de entrada está asociado a múltiples etiquetas de clase? Por ejemplo, cuando escuchas una pieza musical, ¿Cómo te sientes? ¿Te sientes feliz, triste, enfadado, o sientes más de una emoción a la vez? Esta problemática en la literatura se denomina clasificación multietiqueta (de las siglas en inglés MLC). [1,2]

El aprendizaje multietiqueta permite modelar de forma realista este problema, ya que se puede asignar a una pieza musical más de una emoción a la misma vez. Los objetos multietiqueta son habituales en muchas aplicaciones, como la bioinformática, la anotación multimedia, la categorización de textos, y el diagnóstico médico. [3,4] En estos casos, el objetivo del aprendizaje es aprender una función que pueda predecir un subconjunto de etiquetas para un objeto no visto a partir de un conjunto dado de etiquetas.

Dado que el aprendizaje multietiqueta se encuentra todavía en una fase temprana de desarrollo en comparación con otras técnicas de aprendizaje, algunas técnicas actualmente disponibles para otros tipos de aprendizaje no se han desarrollado para este caso específico. Este trabajo pretende resolver esto, y así dotar a la comunidad científica de novedosas técnicas de clasificación multietiqueta.

MÉTODOS

Las principales contribuciones se dirigen a desarrollar medidas de calidad de datos, técnicas de preprocesamiento y predicción, y a mejorar la interpretabilidad de los clasificadores neuronales multietiqueta. Estas se describen brevemente en las siguientes subsecciones.

Medidas de calidad de los datos multietiqueta

El primer resultado de esta investigación son 3 medidas de calidad de datos multietiqueta (MCQ_D, MCQ_E, y MCQ_F), [5] que permiten estimar la complejidad de un problema sin ejecutar ningún algoritmo de aprendizaje MLC. Estas medidas se basan en el cálculo de la consistencia de los datos. Para ello, se utiliza la Teoría de Conjuntos Aproximados (de las siglas en inglés RST), [6] que es probablemente el enfoque más adecuado para analizar la consistencia de los datos. Para derivar estas medidas, se adapta el concepto de consistencia al entorno de las etiquetas múltiples. Las 3 medidas propuestas establecen una relación entre las clases de similitud de los objetos (granulación por condición) y sus clases de decisión (granulación por decisión). Su diferencia radica en que: mientras MCQ_D depende del grado de similitud entre los gránulos por condición y decisión; MCQ_E depende de la similitud entre los rankings por condición y decisión de cada objeto; y MCQ_F de la similitud entre los conjuntos difusos por condición y decisión de cada objeto.

Técnicas de preprocesamiento

Se proponen 3 métodos para la edición de conjuntos de entrenamiento multietiqueta. [7] Estos métodos se basan en las aproximaciones inferior y superior calculadas en la RST para determinar un grado de granularidad adecuado en el conjunto de entrenamiento. El primer método construye un conjunto de entrenamiento como la unión de las aproximaciones inferiores de cada clase de decisión. El segundo método incluye también los objetos de la región frontera, que han sido reetiquetados teniendo en cuenta el grado de pertenencia a cada clase de decisión. El tercer método es similar al segundo, pero omite la conexión entre las clases de decisión para que las etiquetas se traten de forma independiente.

Se desarrollan 3 métodos para generar prototipos en conjuntos de datos multietiqueta independientes del paradigma de aprendizaje. [8,9] Los métodos propuestos utilizan diferentes enfoques de granulación para derivar objetos representativos que sustituyan al conjunto de entrenamiento original. En los 2 primeros métodos propuestos, la granulación del universo se realiza utilizando una relación de similitud que construye clases de similitud (gránulos) de objetos en el universo a partir de los atributos condicionales. El tercer método realiza una granulación del universo a partir de una relación de equivalencia y considerando las diferentes etiquetas existentes en el universo de discurso. Se construye una clase de equivalencia (gránulo) por cada etiqueta y se genera un prototipo por cada gránulo.

Por otro lado, se propone un método que mide el grado de anomalía de un objeto en un conjunto de datos multietiqueta. [10] Esta puntuación o medida cuantifica el grado de irregularidad de un objeto respecto al conjunto de datos. El método se basa en la definición de anomalía dada por Barnett y Lewis. [11] Ellos definen un valor atípico (*outlier*) como una observación (o subconjunto de observaciones) que parece inconsistente con el resto del conjunto de datos. Esta idea se puede modelar utilizando el enfoque de la RST, en el que la consistencia de un objeto se define a partir de la relación entre sus partes predictiva y de decisión. En otras palabras, si la clase de similitud del objeto (es decir, los objetos que son similares a él teniendo en cuenta sus características predictivas) y su clase de equivalencia (es decir, los objetos que son idénticos a él teniendo en cuenta sus etiquetas) son similares, podría decirse que es consistente con respecto al resto de los objetos del conjunto de datos. El grado en que un objeto es anómalo podría depender de cómo el objeto satisface esta relación. El grado asignado a cada objeto estará en el intervalo , donde 0 denota un objeto regular (*inlier*), mientras que 1 indica una fuerte anomalía (*outlier*).

Técnicas de predicción

Se propone una arquitectura neuronal bidireccional para extraer características y etiquetas de alto nivel en problemas de MLC. [12,13] La primera capa de agrupación comprende neuronas que denotan las características y las etiquetas del problema, mientras que, en las capas de agrupación más profundas, las neuronas denotan características y etiquetas de alto nivel que se extraen durante el proceso de construcción. Cada capa de agrupación utiliza una función que detecta los pares de neuronas altamente asociadas y realiza una operación de agregación para obtener las neuronas agrupadas. Para estimar el grado de asociación entre 2 neuronas, utilizamos (a) la correlación de Pearson y (b) la entropía en los gránulos de información [14] que se generan a partir de 2 características o etiquetas. Una vez que las características y etiquetas de alto nivel se han extraído del conjunto de datos, se conectan con 1 o varias capas de procesamiento ocultas que confieren la capacidad de predicción del sistema neuronal. Por último, se realiza un proceso de decodificación [15] para conectar las etiquetas de alto nivel con las originales a través de 1 o varias capas de procesamiento ocultas.

Por otro lado, se presenta un sistema neuronal para resolver problemas de MLC que puedan involucrar características dispersas. [16] La arquitectura de este modelo implica 3 bloques neuronales conectados secuencialmente. El primer bloque consiste en una red multicapa que extrae características de alto nivel. Este bloque reduce la dimensionalidad del espacio codificando la información relevante en las características de alto nivel. El segundo bloque consiste en una *Long-Term*

Cognitive Network (LTCN), [17] que realiza el razonamiento sobre las características extraídas. Por último, el tercer bloque adapta las salidas del bloque recurrente al espacio de etiquetas.

Técnicas de inteligencia artificial explicable

Uno de los retos más importantes de la inteligencia artificial (IA) es la construcción de modelos computacionales eficaces e interpretables, lo que ha dado lugar a la llamada IA explicable (de las siglas en inglés XAI). [18] Si un sistema inteligente resultante de un proceso de aprendizaje automático es capaz de resolver un problema y explicar su solución, la confianza de sus usuarios aumenta, lo que contribuye a la credibilidad de la IA. Esto puede lograrse desarrollando modelos más transparentes o incluyendo una etapa de interpretabilidad post-hoc. El método Layer-wise Relevance Propagation (LRP) es un ejemplo de esta última. Este método proporciona explicaciones en forma de relevancia del espacio de entrada para entender las decisiones de clasificación de las redes neuronales. Se propone una adaptación del método LRP para mejorar la interpretación de los resultados obtenidos por una red neuronal multietiqueta. Para ello, se presenta un enfoque que redistribuye los valores de activación asociados a cada etiqueta hacia los valores de entrada. Este realiza un proceso de agregación de los grados de activación de las etiquetas inferidas, dando como resultado una etiqueta granular a partir de la cual se inicia el proceso de redistribución.

RESULTADOS

Medidas de calidad de los datos multietiqueta

Es lógico pensar que un conjunto de datos con alta consistencia dará lugar a bajos errores en su clasificación. En base a esto se realiza una prueba de correlación entre el valor de consistencia obtenido por las medidas propuestas y el error de clasificación obtenido por el método ML-*k*NN. [19] Para esto, se utiliza la medida de error *Hamming Loss* (HL), [1] que es una de las más usadas en la literatura MLC para medir el rendimiento de los clasificadores. Además, se emplean varios conjuntos de datos multietiqueta tomados de los repositorios MULAN [20] y RUMDR. [21] Los resultados alcanzados muestran una fuerte correlación negativa entre los valores de consistencia obtenidos por las 3 medidas propuestas y el rendimiento del algoritmo, lo cual confirma la hipótesis.

Técnicas de preprocesamiento

Con el fin de explorar el rendimiento global de los métodos propuestos se utiliza el algoritmo ML-kNN, [19] mientras que se emplea la medida HL [1] como medida de evaluación del rendimiento. Las simulaciones numéricas mostraron que los métodos de reducción (de edición y generación de prototipos) propuestos permiten una reducción de hasta el 80 % del número de instancias en algunos de los conjuntos de entrenamiento utilizados. De hecho, en algunos casos se observa un aumento del poder discriminatorio del algoritmo ML-kNN.

Por otra parte, el rendimiento del método de detección propuesto se comparó con 2 algoritmos reportados en la literatura: Exact k-Nearest Neighbor Score y Average k-Nearest Neighbor Score. [22] Ambos algoritmos fueron adaptados a la problemática multietiqueta, y fueron seleccionados porque también proporcionan un grado de anomalía para cada objeto. Los resultados mostraron que el método propuesto es más eficaz que los otros 2 métodos en distinguir entre un outlier y un inlier. La razón de esto es porque estos métodos no tienen en cuenta la relación existente entre las características predictivas y las etiquetas de un objeto.

Técnicas de predicción

Las simulaciones numéricas sobre la arquitectura neuronal bidireccional propuesta demostraron su capacidad de reducción en términos del número de parámetros en las redes neuronales profundas. Además, este poder de reducción se logra sin perjudicar su poder discriminatorio. Sin embargo, la primera variante pensada para la función de asociación (la variante basada en la correlación de Pearson) requiere que las características y las etiquetas tengan un cierto grado de correlación, lo que puede no ser aplicable en todas las situaciones. En este caso, el segundo enfoque propuesto (la variante basada en la entropía de los gránulos de información), reportó mayores valores de reducción en conjuntos de datos con bajos valores de correlación entre sus características y etiquetas.

Por otro lado, las simulaciones numéricas sobre la segunda propuesta neuronal sugieren que el modelo propuesto se comporta bien cuando se añaden más capas abstractas al segundo bloque mientras se utilizan valores de tasa de aprendizaje bastante pequeños. Además, la comparación con otros métodos MLC reportados en la literatura [16] muestra que el modelo propuesto es un buen competidor en términos de HL. [1]

Técnicas de inteligencia artificial explicable

La eficacia del enfoque de interpretación *post-hoc* propuesto se evalúa siguiendo los 2 criterios de evaluación existentes en Montavon G *et al*, 2018, [23] demostrándose la calidad de las explicaciones obtenidas por este.

El método propuesto se aplica en la explicación de la salida de una red neuronal multietiqueta que detecta coinfec-

ciones secundarias en pacientes infectados con SARS-CoV-2. Las coinfecciones asociadas a la infección SARS-CoV-2 se clasifican en infecciones bacterianas e infecciones micóticas, [24] un paciente puede desarrollar una, ambas o ninguna. La inclusión de un enfoque de interpretación *post-hoc* permitió identificar las variables de entrada que influyen en que un paciente esté coinfectado con 1 o más de 2 infecciones simultáneamente, lo cual permitió identificar patrones que aportan nuevo conocimiento a los especialistas.

DISCUSIÓN

Se aportan novedosas medidas para estudiar la calidad de los datos, métodos para mejorar la calidad de estos, métodos para construir sistemas inteligentes mediante un proceso de aprendizaje a partir de los datos, y un método para explicar las respuestas de estos sistemas multietiqueta. Todas estas propuestas son técnicas efectivas (en términos de eficiencia y eficacia) y novedosas dentro del campo de la MLC, ya que hasta su desarrollo eran insuficientes para la comunidad científica dedicada a este campo del aprendizaje.

La actualidad de esta propuesta se evidencia en su aplicación a un problema real como es el caso de la detección temprana de coinfecciones secundarias en pacientes infectados con SARS-CoV-2.

Conclusiones

La investigación desarrollada aborda el problema de la MLC desde los datos hasta la explicación de la respuesta de un sistema neuronal al procesar los datos. Teniendo en cuenta que el objetivo era estudiar el problema del MLC de la forma más exhaustiva posible, se hacen aportaciones en cada etapa del proceso de descubrimiento de conocimiento.

REFERENCIAS BIBLIOGRÁFICAS

1. Herrera F, Charte F, Rivera AJ, Del Jesus MJ. Multilabel classification. Multilabel Classification: Springer; 2016:17-31 p.
2. Gibaja E, Ventura S. Multi-label learning: a review of the state of the art and ongoing research. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2014;4(6):411-44.
3. Gutierrez BJ, Zeng J, Zhang D, Zhang P, Su Y. Document classification for covid-19 literature. arXiv preprint arXiv:200613816. 2020.
4. Yu Q, Wang J, Zhang S, Gong Y, Zhao J. Combining local and global hypotheses in deep neural network for multi-label image classification. Neurocomputing. 2017;235:38-45.
5. Bello M, Nápoles G, Vanhoof K, Bello R. Data quality measures based on granular computing for multi-label classification. Information Sciences. 2021;560:51-67.
6. Pawlak Z. Rough sets. International journal of computer & information sciences. 1982;11(5):341-56.
7. Bello M, Nápoles G, Vanhoof K, Bello R, editors. Methods to edit multi-label training sets using rough sets theory. International Joint Conference on Rough Sets; 2019: Springer.
8. Bello M, Nápoles G, Vanhoof K, Bello R. On the generation of multi-label prototypes. Intelligent Data Analysis. 2020;24(S1):167-83.
9. Bello M, Nápoles G, Vanhoof K, Bello R, editors. Prototypes Generation from Multi-label Datasets Based on Granular Computing. Iberoamerican Congress on Pattern Recognition; 2019: Springer.
10. Bello M, Nápoles G, Morera R, Vanhoof K, Bello R, editors. Outliers Detection in Multi-label Datasets. Mexican International Conference on Artificial Intelligence; 2020: Springer.
11. Barnett V, Lewis T. Outliers in statistical data. osd. 1984.
12. Bello M, Nápoles G, Sánchez R, Bello R, Vanhoof K. Deep neural network to extract high-level features and labels in multi-label classification problems. Neurocomputing. 2020;413:259-70.
13. Bello M, Nápoles G, Sánchez R, Vanhoof K, Bello R, editors. Feature and label association based on granulation entropy for deep neural networks. International Joint Conference on Rough Sets; 2020: Springer.
14. Yao Y. Probabilistic approaches to rough sets. Expert systems. 2003;20(5):287-97.
15. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. science. 2006;313(5786):504-7.
16. Nápoles G, Bello M, Salgueiro Y. Long-term Cognitive Network-based architecture for multi-label classification. Neural Networks. 2021;140:39-48.
17. Nápoles G, Vanhoenshoven F, Falcon R, Vanhoof K. Nonsynaptic error backpropagation in long-term cognitive networks. IEEE transactions on neural networks and learning systems. 2019;31(3):865-75.
18. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. 2020;58:82-115.
19. Zhang M-L, Zhou Z-H. ML-KNN: A lazy learning approach to multi-label learning. Pattern recognition. 2007;40(7):2038-48.
20. Tsoumakas G, Xioufis E, Vilcek J, Vlahavas I. MULAN multi-label dataset repository. URL http://mulan.sourceforge.net/datasets.html. 2014.
21. Charte F, Charte D, Rivera A, del Jesus MJ, Herrera F, editors. R ultimate multilabel dataset repository. International Conference on Hybrid Artificial Intelligence Systems; 2016: Springer.
22. Aggarwal CC, editor Outlier analysis. Data mining; 2015: Springer.
23. Montavon G, Samek W, Müller K-R. Methods for interpreting and understanding deep neural networks. Digital Signal Processing. 2018;73:1-15.
24. Aguilera Calzadilla Y, Díaz Morales Y, Ortiz Díaz LA, Gonzalez Martínez OL, Lovelle Enríquez OA, Sánchez Álvarez MdL. Infecciones bacterianas asociadas a la COVID-19 en pacientes de una unidad de cuidados intensivos. Revista Cubana de Medicina Militar. 2020;49(3).

Recibido: 06/10/2022
Aprobado: 13/01/2023

Agradecimientos
Los autores agradecen a todos aquellos que colaboraron de una forma u otra en la presente investigación: Yamisleydi Salgueiro Sicilia, Ricardo Sánchez Alba, Rafael Alejandro Fernández Morera, Leticia Arco García, y Bárbara Toledo Pimentel.

Conflictos de intereses
Los autores declaran que no hay conflicto de intereses con ninguna entidad o autor.

Contribuciones de los autores
Conceptualización: Rafael E. Bello Pérez, Gonzalo Nápoles
Curación de datos: Marilyn Bello García, Yaumara Aguilera Calzadilla
Análisis formal: Rafael E. Bello Pérez, Gonzalo Nápoles, Marilyn Bello García
Investigación: Marilyn Bello García
Metodología: María M. García Lorenzo
Recursos: Koen Vanhoof, Gonzalo Nápoles
Software: Marilyn Bello García
Supervisión: Rafael E. Bello Pérez, Gonzalo Nápoles, Koen Vanhoof
Validación: Yaumara Aguilera Calzadilla
Redacción-borrador original: Marilyn Bello García
Redacción-revisión y edición: Marilyn Bello García, Rafael E. Bello Pérez, Gonzalo Nápoles, Koen Vanhoof, María M. García Lorenzo

Financiamiento
La investigación ha estado parcialmente financiada por la Universidad de Hasselt, Bélgica, mediante becas de intercambio académico.

Cómo citar este artículo
Bello García M, Bello Pérez RE, Nápoles G, Vanhoof K et al. Desarrollo de técnicas para el preprocesamiento y la predicción de problemas de clasificación multietiqueta. An Acad Cienc Cuba [internet] 2023 [citado en día, mes y año];13(3):e1344. Disponible en: http://www.revistaccuba.cu/index.php/revacc/article/view/1344