UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

APRENDIZAJE MULTIETIQUETA DE PATRONES GEOMÉTRICOS EN OBJETOS DE HERENCIA CULTURAL

PROPUESTA DE TESIS PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN COMPUTACIÓN Y MAGÍSTER EN CIENCIA DE DATOS

MATÍAS VERGARA SILVA

.....

PROFESOR GUÍA:
BENJAMÍN BUSTOS CÁRDENAS

.....

PROFESOR COGUÍA:
IVÁN SIPIRÁN MENDOZA

SANTIAGO DE CHILE
2022

1. Introducción

El presente documento versa sobre la propuesta de tesis de Matías Vergara, estudiante de Ingeniería Civil en Computación y del Magíster en Ciencia de Datos, ambos programas de la Universidad de Chile. El tema propuesto lleva por nombre "Aprendizaje multietiqueta de patrones geométricos en objetos de herencia cultural" y se desarrolla con los profesores Benjamín Bustos e Iván Sipirán como guía y coguía, respectivamente.

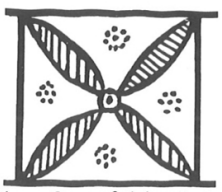
A fin de entender el problema a tratar, resulta primero necesario adentrarse en lo que es la *herencia cultural*. Definida como aquel patrimonio de bienes tangibles o intangibles de un grupo o sociedad que se hereda de generaciones pasadas [16], el estudio y conservación de herencia cultural es un área presente en ciencias tales como la arqueología, la paleontología y la antropología, entre muchas otras.

Existe sin embargo un término acuñado para referirse específicamente a los bienes tangibles. Se trata de los *objetos de herencia cultural*, definidos como cualquier bien mueble de importancia cultural que requiera protección - trátese de artefactos arqueológicos, especímenes paleontológicos/geológicos, meteoritos o cualquier otro objeto con un significado cultural [14] -. Dichos bienes son materia importante de conservación y catalogación, en cuanto su estudio permite alcanzar una mayor comprensión de la cultura que les subyace.

Tales objetos se hacen particularmente presentes en la Arqueología, ciencia que busca estudiar la diversidad humana a través del registro material y que, como consecuencia, da lugar a numerosas excavaciones e incursiones en búsqueda de vestigios de culturas pasadas, encontrándose así con múltiples objetos los cuales rescatan, describen y, una vez extraída toda la información posible, ponen a disposición de instituciones de conservación.

En todo el proceso anterior la etapa de descripción es probablemente la más compleja, dado que involucra múltiples tareas según el tipo de objeto, su origen y la característica en estudio. Un ejemplo cercano de dicha complejidad se encuentra en el caso de los múltiples esfuerzos realizados por construir una tipología y seriación de la cerámica de la cultura Chimú [3], estado andino que se estableciese en el actual Perú entre los siglos X y XV.

Más aún, existe un caso particular del proceso descriptivo el cual resulta sumamente desafiante y en el cual se centra esta investigación. Se trata del etiquetado de patrones geométricos, tarea en la cual un patrón es asociado a múltiples "etiquetas" - normalmente sustantivos o adjetivos, o una combinación de ambos - que responden a las distintas propiedades geométricas del mismo. Un ejemplo de ello se presenta en la Figura 1, donde el patrón (arriba) es asociado a las etiquetas *quatrefoil, hatched, with central dotted circle, metopal* (abajo).



16e Quatrefoil, hatched, with central dotted circle, metopal

Figura 1: Patrón 16e de *Geometrischer Vasen: Ein Kompendium* [11] y sus etiquetas.

De esta manera, el problema que se busca resolver consiste en la inexistencia de una herramienta de apoyo al proceso, la cual lo haga más simple al mismo tiempo que introduzca cierto estandarizado de etiquetas. Las razones por las cuales esto representa un problema son, por un lado, lo tediosa que la tarea resulta dada su alta complejidad y la consecuente necesidad de invertir una gran cantidad de horas hombres en su desarrollo, y por otro lado, la falta de acuerdo entre expertos con respecto a qué etiquetas utilizar y en qué ocasiones. Ambos aspectos se detallarán a continuación.

En primer lugar, la complejidad del proceso se debe principalmente a la gran cantidad de información geométrica que un mismo patrón puede contener, dando así lugar a un basto universo de etiquetas posibles. Esto se traduce en un trabajo extensivo y eventualmente incompleto, en el cual los expertos pueden invertir horas tratando de asignar todas las etiquetas necesarias y aún así no lograr reflejar todas las características importantes del patrón.

En segundo lugar, la falta de acuerdo entre etiquetadores encuentra sus orígenes en una diferencia de trasfondo, lo cual a menudo da lugar a variaciones en la consistencia de los datos [12]. Por ejemplo, un patrón podría recibir las etiquetas *ave*, *vertical* y *metopa* por un experto A, mientras que un experto B podría asignarle las etiquetas *pájaro* y *vertical*, evidenciando así dos problemas: el uso de sinónimos para referirse a la misma característica (*pájaro* y *ave*), y la ausencia de la etiqueta *metopa* en el etiquetado de B.

Ante las razones descritas surge entonces la pregunta de por qué realizar un etiquetado correcto resulta tan importante. La respuesta se encuentra en el hecho de que el etiquetado refleja una clasificación tipológica en la cual la cantidad de objetos individuales con un cierto contexto arqueológico resulta esencial para deducir aspectos socioculturales, económicos o conductuales de dichos contextos [1], por lo cual un etiquetado completo y sin ambigüedad es más indicativo del significado que el patrón en estudio pudo tener en el pasado, así como de la utilidad del objeto que lo contiene. Ambos resultados (significado del patrón y utilidad del objeto) contribuyen finalmente a aumentar el entendimiento de la cultura subyacente, facilitando además la *musealización* del objeto.

En este sentido, lo que se busca es desarrollar una herramienta que sirva de apoyo al experto mediante la sugerencia de etiquetas relevantes con nombres estandarizados, atacando así los dos problemas detectados: la complejidad del problema, al reducir la cantidad de etiquetas que deben ser ideadas por el experto, y la falta de acuerdo, al sugerir etiquetas estandarizadas.

Una vez descrito el problema, su contexto y el valor que una eventual solución podría aportar al área, queda aún pendiente aclarar cómo todo esto puede ser abarcado desde un enfoque de Ingeniería Civil en Computación y de Ciencia de Datos. Para dar respuesta a esta nueva interrogante, es primero necesario introducir brevemente lo que es la tarea de *clasificación* en *aprendizaje supervisado.*

Una necesidad frecuente en Ciencia de Datos es la de una herramienta que, al recibir cierta entrada, responda asignando una clase l dentro de un conjunto L de clases tal que $|L| > 1$. Esta tarea recibe el nombre de "clasificación" y acostumbra ser aproximada desde un enfoque de aprendizaje supervisado, técnica en la cual se "entrenan" algoritmos y modelos computacionales para aprender a entregar la clase correcta a partir de observar un conjunto de

datos cuya clase es conocida y posteriormente usar dicha información para intentar predecir la clase de ejemplos nuevos, en un proceso de prueba y error. Existen múltiples variaciones de dicha tarea, siendo ejemplo quizá más clásico aquel en donde $|L| = 2$, caso en el que recibe el nombre de *clasificación binaria*. Un caso más complejo es el de la *clasificación multiclase*, que se halla cuando $|L| > 2$. En ambos casos, una característica se mantiene invariante: cada ejemplo pertenece y es asignado a una única clase, tratándose así de clasificación *singular* [20].

Existe sin embargo una familia de problemas hermanos a la clasificación singular, menos explorados y ampliamente desafiantes: se trata de la clasificación *multietiqueta*, donde en lugar de asociar cada entrada a una clase $l \in L$, se asocia a un conjunto de clases $Y \subseteq L$. Un ejemplo de ello sería la clasificación de enfermedades a partir de diagnósticos médicos, donde un mismo paciente podría estar sufriendo de diabetes y cáncer simultáneamente [20]. Otro ejemplo se encuentra en la genética, donde un mismo gen puede ser responsable de codificar múltiples características.

Es a través de este tipo de clasificación multietiqueta, más desafiante pero también más cercana al mundo real, que la presente investigación busca dar lugar a una herramienta de apoyo al etiquetado de patrones, tras identificar la evidente relación entre el proceso de asignar múltiples etiquetas a un patrón y el aprendizaje multietiqueta. Se pretende que, tras una selección y tratamiento metódico de datos, se logre llegar a un conjunto de entradas que permita, a través de las bondades de la programación, entrenar distintos algoritmos y/o modelos de clasificación multietiqueta sobre los patrones geométricos, logrando así - en un panorama optimista - la automatización del etiquetado.

Ahora bien, existen dos variables cuyo comportamiento *a priori* es desconocido y que podrían incidir ampliamente en si se alcanza o no la automatización. La primera de ellas guarda relación con la factibilidad de obtener un conjunto de datos lo suficientemente grande y representativo, variable la cual está restringida por trabajo previo de los profesores guía a una única fuente de datos (que se discutirá más adelante). La otra variable guarda relación con la existencia de una taxonomía real en las etiquetas, pues son este tipo de características - agrupaciones ordenadas, construidas en base a relaciones " naturales" - las más factibles de aprender mediante el enfoque propuesto.

Pese a lo anterior, existe una razón adicional para llevar a cabo la investigación aún sin tener certeza de dichas variables. Se trata de la inexistencia a la fecha de una solución similar, lo cual tiene por efecto el que no sea necesario alcanzar las sugerencias perfectas o el automatizado total del proceso para que signifique un aporte significativo al área, si no que por el contrario, el solo hecho de desarrollar una herramienta de apoyo - que sugiera un conjunto de etiquetas a evaluar y/o corregir por el experto - significaría un avance enorme.

De esta manera, la meta del trabajo a realizar se define como el concebir una herramienta de apoyo al etiquetado de patrones geométricos, la cual actúe como un *sistema recomendador* que permita hacer del etiquetado un proceso más eficiente y eficaz mediante la sugerencia de etiquetas razonables al experto y esperando que ello siente las bases para, en un futuro, alcanzar la automatización total.

2. Estado del Arte

Del etiquetado en Arqueología

En la actualidad, el proceso del etiquetado de patrones se lleva a cabo de forma completamente manual: el arqueólogo accede a un objeto sin etiquetar (generalmente a partir de excavaciones), lo estudia en búsqueda de patrones interesantes y, de encontrarlos, realiza un *sketch* de ellos para posteriormente dedicarse a la asignación de etiquetas, lo cual puede tomar desde minutos hasta horas dependiendo de la complejidad de los patrones y la motivación que el objeto signifique para el etiquetador. En todo el transcurso no existe ninguna herramienta de apoyo al experto.

Sin embargo, existen otros procesos propios de la arqueología donde la computación ha logrado contribuir con herramientas de apoyo, especialmente en lo que a computación gráfica y el análisis de formas tridimensionales se refiere - área que ha presentado una simbiosis histórica con el dominio de los objetos culturales [17] -. Ejemplos de ello se encuentran en la digitalización 3D, la reconstrucción y restauración visual de objetos, el desarrollo de interfaces inmersivas, entre otros.

Un ejemplo quizá más cercano al tema de esta propuesta se encuentra en el etiquetado de estilos de vasijas de herencia cultural, donde se desarrollase una herramienta que buscaría, a partir de un enfoque de *gamificación*, reducir la probabilidad de observar vasijas con etiquetados incompletos al aumentar la motivación del experto mediante un juego. Si bien dicha herramienta habría logrado cumplir su objetivo, el estudio de sus resultados indicaría que también habría aumentado el nivel de desacuerdo entre expertos [12]. La Figura 2 muestra una captura de dicho sistema.

Figura 2: Sistema de etiquetado de estilos de vasijas mediante gamificación.

Es así como, pese a lo inexplorado del etiquetado de patrones geométricos desde un enfoque de aprendizaje de maquinas, el potencial que guarda incorporar herramientas computacionales en los procesos arqueológicos es bien conocido para el área, así como lo es también el problema del etiquetado incompleto o inconsistente, el cual resulta además transversal a los múltiples catálogos posibles (de estilo, de forma, de patrones, entre otros).

En el caso particular del etiquetado de patrones geométricos, dicho problema ha servido de motivación para que los expertos busquen, a través de su obra, sentar una taxonomía base

sobre la cual sus colegas puedan guiarse. Uno de los intentos más famosos es el libro *Geometric Vasen: Ein Kompendium* [11], autoría del prof. Dr. Norbert Kunisch, arqueólogo de renombre mundial que buscase establecer una taxonomía de etiquetas a través de ejemplos provenientes de excavaciones en Grecia durante el siglo XX. Dicho libro lograría reconocimiento en el área, más no habría alcanzado a establecerse como una referencia del etiquetado. Una posible razón para ello se encuentra en la misma naturaleza tediosa del proceso, donde el incorporar una tarea adicional - tal como la revisión de la obra de un tercero a modo de guía - no hace más que disminuir la motivación del experto por alcanzar un trabajo completo [12].

De esta manera, la investigación propuesta se plantea como una innovación sin precedentes para la el etiquetado de patrones geométricos, pero con el amplio respaldo de aplicaciones similares en otras áreas de la arqueología. El problema que abarca, por otro lado, se muestra como una dificultad ampliamente conocida y con un historial de intentos de solución sin éxito, lo cual refleja a la vez la importancia de encontrar una solución y la alta complejidad del mismo.

Del aprendizaje multietiqueta

Una vez discutido el estado del arte del etiquetado, queda pendiente aún estudiar el del enfoque propuesto. En este sentido, vale la pena dividir el estudio en tres secciones: una para presentar en mayor profundidad los desafíos del aprendizaje multietiqueta, y otras dos para describir los dos acercamientos más comunes: desde aprendizaje de máquinas tradicional y desde aprendizaje profundo.

Aprendizaje multietiqueta y sus desafíos

En la actualidad, los métodos de clasificación multietiqueta (MLC, por sus siglas en inglés) son cada vez más solicitados, en cuanto el escenario al que buscan responder - aquel donde los ejemplos se relacionan con un conjunto de etiquetas $Y \subseteq L$ - se hace presente en múltiples aplicaciones modernas, tales como la clasificación de funciones de proteínas, la categorización de música y la clasificación semántica de escenas [19]. En muchos casos resulta incluso más natural que la clasificación singular, puesto que, por ejemplo, una foto de animales domésticos - ejemplo por excelencia de algoritmos de clasificación singular - podría no solo contener un *gato* o un *perro*, si no que también *arboles*, *nubes* y *pasto*. Naturalmente, extraer toda esta información resultaría mucho más útil que solo indicar la especie protagonista de la imagen.

Al tratarse de una tarea que contiene a la clasificación singular, la clasificación multietiqueta mantiene sus desafíos e incorpora otros adicionales. Por un lado, mantiene aquellos relacionados con la eficiencia y rendimiento: la necesidad, en primer lugar, de clasificadores que no escalen en tiempo o en complejidad espacial con el número de etiquetas y, por otro lado, de alcanzar resultados eficaces y con una alta capacidad de generalización. Incorpora, sin embargo, un desafío usualmente resuelto para la clasificación singular: la evaluación de resultados, cuya dificultad radica en el hecho de que las métricas para clasificación singular dan lugar frecuentemente a evaluaciones demasiado estrictas en el caso múltiple.

Es esta última dificultad la que ha servido como motivación para el desarrollo de distintas métricas específicas para MLC, las cuales en general se pueden categorizar en dos grandes grupos: aquellas métricas basadas en ejemplos [6][7] y aquellas basadas en etiquetas [22].

Las métricas basadas en los ejemplos, por un lado, trabajan evaluando la eficacia del sistema de aprendizaje en cada ejemplo de prueba por separado, y luego retornando el promedio obtenido en el conjunto de prueba. Ejemplos de este tipo de métricas se encuentran en el *Hamming Score* (también conocido como *subset accuracy* [5]), *Hamming Loss* y la versión promediada por ejemplo de las métricas clásicas (*precision, recall, F-score,* etc).

A modo de ejemplo, las Ecuaciones 1 y 2 presentan la formulación matemática del *Hamming Score* y la versión basada en ejemplos de *Recall*, donde D corresponde al conjunto de ejemplos e Y_i y Z_i a los conjuntos de etiquetas reales y predichas para un ejemplo i, respectivamente. Ambas métricas cobrarán especial relevancia en secciones próximas.

$$\text{Hamming-Score} = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|} \tag{1}$$

$$\text{Recall} = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \cap Z_i|}{|Y_i|} \tag{2}$$

Las métricas basadas en etiquetas, por otro lado, trabajan evaluando la eficacia del sistema en cada etiqueta por separado, y luego retornando el promedio macro o micro sobre todas las etiquetas. En este caso, las métricas singulares trabajan bien, pero la selección de promedio micro/macro debe hacerse con cuidado, en especial en casos con un alto desbalance de etiquetas.

Lo anterior podría llevar a pensar con justa razón que uno de los desafíos al momento de realizar una clasificación multietiqueta es entonces la elección de métricas adecuadas. Lo anterior es correcto, y sin embargo, no es el desafío principal al momento de abarcar dichas tareas, pues ese título le pertenece a una dificultad distinta: se trata de qué tantas etiquetas se busca predecir y el nivel de presencia de estas en los datos.

Al respecto, se han realizado esfuerzos para establecer una metodología que permita responder a la pregunta *"¿Qué tan multietiqueta es un dataset?"* [19]. Dichos esfuerzos han dado lugar a dos métricas específicas para la descripción de conjuntos de datos: se trata de *label average* y *label density*, cuyas formulaciones matemáticas se presentan en las Ecuaciones 3 y 4, respectivamente.

$$\text{Label Average (D)} = \frac{1}{|D|}\sum_{i=1}^{|D|}|Y_i| \tag{3}$$

$$\text{Label Density (D)} = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i|}{|L|} \tag{4}$$

Mientras por un lado *label average* refleja la cantidad de etiquetas promedio por ejemplo, *label density* complementa esta información dividiendo por la cantidad total de etiquetas,

entregando así una información adicional: qué tantos casos de etiquetas "positivas" hay en los datos (entendiendo por positivo el caso donde la asignación de la etiqueta para un ejemplo es correcta, y por negativo cuando una etiqueta no corresponde), entregando así una cierta noción de desbalance de casos negativos/positivos promedio por etiqueta. Dicha métrica, que toma un valor 1 si todas las entradas se relacionan con todas las etiquetas y 0 si todas las entradas se relacionan con ninguna etiqueta, resulta fundamental al momento de evaluar un conjunto de datos, en cuanto tiene una incidencia directa en la capacidad de aprendizaje de los algoritmos [2].

Recientemente se ha acuñado además un término para referirse específicamente a los problemas de clasificación multietiqueta donde el número de etiquetas a predecir es del orden de millones: se trata de la *clasificación multietiqueta extrema* [13] o *XMLC*, por sus siglas en inglés. Dicho caso, que forma parte de los problemas conocidos como *Challenging MLC* - junto al aprendizaje con etiquetas incompletas (*Explicit MLC*) y erróneas (*Implicit MLC)* - incorpora sin embargo dos dificultades adicionales: por un lado, un aprendizaje computacionalmente muy costoso, en especial con los modelos más tradicionales (que se verán a continuación) y, por el otro, el hecho de necesitar cientos de miles, o incluso millones de ejemplos de entrenamiento etiquetados [13].

Existe además un último desafío propio del aprendizaje multietiqueta y que guarda estrecha relación con esta noción de desbalance que brindaría el estudio de la *label density*. Se trata de la dificultad presente al momento de intentar balancear los datos, en cuanto el hecho de que las etiquetas vengan en conjuntos conlleva a que el balance de una puede resultar en el desbalance de otra. Lo anterior descarta de inmediato métodos tales como el sampleo estratificado, y ha sido tratado por la literatura principalmente a través de técnicas de asignación de pesos y del uso de funciones de pérdida diseñadas específicamente para lidiar con este aspecto [9].

Una vez descritos los principales desafíos del multietiquetado y las soluciones propuestas en el estado del arte, se procede a describir las técnicas más frecuentes en sus dos enfoques: el enfoque tradicional y el de aprendizaje profundo.

Enfoque tradicional

Los métodos de clasificación multietiqueta tradicionales corresponden a aquellos en donde las características de entrada para la clasificación se obtienen a partir de un proceso de *feature engineering* previo, y se pueden dividir en dos grandes categorías: aquellos métodos de transformación del problema y aquellos de adaptación del algoritmo [27].

Los métodos de transformación del problema, por un lado, corresponden al mapeo del problema multietiqueta hacia múltiples problemas singulares (uno por cada etiqueta) a resolver por clasificadores base, conformando así la forma más intuitiva de lidiar con tal desafío. El representante por excelencia de este enfoque es el algoritmo de *Binary Relevance* (BR) [25], el cual consiste en entrenar un clasificador binario para cada etiqueta y predecir con la concatenación de la salida de cada clasificador. Dicho método, aunque poderoso, es también denominado en ocasiones como un método ingenuo, en cuanto realiza una simplificación que puede resultar extrema: la inexistencia de relaciones entre las etiquetas.

Existen sin embargo otros métodos de la misma familia que no requieren de supuestos tan extremos, y que sin embargo presentan otras debilidades. Un ejemplo clásico de ello es *Classifier Chain* (CC) [15], el cual a través de una transformación a múltiples problemas binarios encadenados permite rescatar información relativa a relaciones entre etiquetas, mas bajo un importante costo en complejidad espacial que lo vuelve inapropiado para grandes conjuntos de etiquetas y una fuerte dependencia del orden en que las etiquetas se presenten.

Otras alternativas que pertenecen también a la categoría de transformación del problema se encuentran en los métodos que no transforman el multietiquetado a múltiples problemas binarios si no a un único problema multiclase. En este caso el ejemplo clásico es *Label Powerset* (LP), algoritmo en el cual el objeto de predicción no son las etiquetas por sí solas si no más bien las combinaciones de etiquetas - conformando así un problema multiclase donde cada clase es una combinación-. Dicho enfoque logra captar las relaciones entre etiquetas, mas sufre de dos grandes debilidades: la necesidad de requerir que todas las combinaciones posibles de etiquetas se encuentren en los datos de entrenamiento (lo cual en general es difícil de conseguir) y una alta probabilidad de subajuste en espacios de alta dimensión.

Una versión mejorada de LP se halla en *Random k-Labelsets* (RakelD) [22], método el cual busca hacer frente a las debilidades de dicha técnica mediante la construcción de particiones aleatorias del espacio de etiquetado y su posterior uso como combinaciones para LP, presentando sin embargo una fuerte dependencia con la probabilidad de observar las particiones aleatorias en los datos reales. *Ovelapped Random k-Labelsets* (RakelO), por otro lado, busca atacar este último punto al convertir las particiones en subconjuntos potencialmente sobrepuestos.

En la otra vereda se encuentran los métodos de adaptación del algoritmo, los cuales buscan extender algoritmos de clasificación singular para tratar con la clasificación multietiqueta. Ejemplos de métodos pertenecientes a esta categoría son *Multilabel k-Nearest Neighbors* (MLkNN) [27] y *Binary Relevance k-Nearest Neighbors* (BRkNN) [23], los cuales mantienen una idea general de ampliar el funcionamiento clásico de *kNN* con algún algoritmo adicional que permita traducir los vecindarios hacia las múltiples salidas esperadas (a través de inferencia bayesiana en el caso de MLkNN y mediante BR en el de BRkNN).

Otro método perteneciente a la clase de adaptación del algoritmo se encuentra en *Multilabel Twin Support Vector Machines* (MLTSVM) [4], el cual busca capturar la información multietiqueta a través de la determinación de múltiples hiperplanos dados por el entrenamiento de un SVM para cada etiqueta.

Nuevamente, cada método tiene sus ventajas y desventajas, siendo una característica transversal la condición de que la distancia entre etiquetas sea un buen predictor para su asignación, además del hecho de descansar por detrás en la transformación del problema mediante clasificadores base.

Existe sin embargo una característica propia del enfoque tradicional que pone en desventaja a ambas familias de métodos. Se trata de la fuerte dependencia que se genera con la calidad de los descriptores o *features* [8], la cual viene determinada por un proceso previo el cual no se retroalimenta de forma automática y que puede ser significativamente demandante en tiempo (tiempo que además es, con alta probabilidad, de expertos). Es esta desventaja la

que motiva a buscar alguna estrategia *End-To-End*, donde la generación de *features* se presente como parte del método mismo y no como un paso previo desconectado del aprendizaje multietiqueta. Es así como el enfoque de aprendizaje profundo comenzaría a ganar relevancia en el dominio de la clasificación multietiqueta.

Enfoque de aprendizaje profundo

En los años recientes, las redes neuronales profundas han presentado una creciente atención debido a su capacidad de aprender automáticamente una representación conveniente de los datos, integrando así el aprendizaje de representaciones y la clasificación en un mismo *framework* End-To-End [8]. Dicha bondad respondería directamente a las desventajas más importantes del enfoque tradicional, por lo cual despertaría la motivación necesaria para dejar este último atrás y pasar a experimentar con arquitecturas más complejas para la clasificación multietiqueta.

Es así como, tras años de investigación, dos grandes familias de arquitecturas se conformarían: por un lado, aquellas que abordan el problema desde un enfoque de BR y que serían conocidas como *Binary Relevance Neural Networks* (BRNN) y, por el otro, aquellas que se adaptarían al problema multietiqueta mediante el uso de una *softmax* como función de salida, recibiendo así el nombre de *Threshold Dependent Neural Networks* (TDNN) [8].

En el caso de las BRNN, su funcionamiento corresponde, en términos generales, a aquel del BR tradicional pero utilizando redes neuronales como clasificadores base. De esta manera, se construye una red neuronal binaria para cada etiqueta y luego se utiliza el conjunto de predicciones independientes como predicción multietiqueta. Desde luego, esto mantiene la principal desventaja de BR: asumir la inexistencia de relaciones entre etiquetas, y suma además un importante costo en tiempo y espacio al entrenar tantas redes profundas como etiquetas. Gana, sin embargo, la posibilidad de contar con descriptores que no solamente han sido generados de forma interna por las redes y entrenados en un proceso de retroalimentación, si no que además son particulares a cada etiqueta (una misma entrada tendrá tantas representaciones potencialmente distintas como etiquetas se busque predecir).

Las TDNN, por otra parte, basan su funcionamiento en construir una única red neuronal y entrenarla para predecir las probabilidades de todas las etiquetas posibles mediante una función *softmax*, donde las probabilidades sumen 1. Posteriormente, un mecanismo adicional de umbral es aplicado para transformar las probabilidades (que hasta ese momento responden a un problema multiclase) en una salida multietiqueta. En este sentido, las TDNN ganan en cuanto a explotación de relaciones y economía de tiempo y espacio se refiere, pero presentan la desventaja de incorporar un mecanismo adicional de *thresholding*, lo cual no solo acaba con la deseada propiedad de ser un método End-To-End, si no que además consiste en un desafío por sí mismo, en cuanto construir una función umbral efectiva es también una tarea llena de desafíos para el aprendizaje multietiqueta [26].

De esta manera, la clave para tratar el problema de clasificación multietiqueta desde un enfoque de aprendizaje profundo pasa principalmente por encontrar aquella arquitectura que logre extraer y entender de la mejor manera posible las características inherentes a las etiquetas desde los patrones. No basta, sin embargo, con tener en consideración solamente las ventajas y desventajas de cada una, pues al tratarse de un enfoque de aprendizaje profundo,

se suman todas las variables propias del mismo: la elección de un optimizador adecuado, el definir una función de pérdida apropiada, entre otros.

Un punto importante a considerar para la investigación es que, si bien las arquitecturas profundas más avanzadas presentan una poderosa capacidad de aprendizaje y por ende más potencial para el problema de clasificación multietiqueta, vienen normalmente acompañadas por una alta complejidad en términos de costos de entrenamiento y predicción. Es por esta razón que la literatura recomienda comenzar explorando el diseño de arquitecturas ligeras que permitan tener un entrenamiento y predicción eficientes, y de esta manera ir buscando progresivamente métodos profundos más avanzados para clasificación multietiqueta [13]. En este sentido, se considera que arquitecturas tales como *Canonical Correlated Autoencoder (C2AE)* [24] y redes *feed-forward* sobre redes recurrentes (*RNN*) representan un buen punto de partida.

3. Objetivos

Objetivo General

Concebir una herramienta computacional que, a partir de un patrón geométrico proveniente de un objeto de herencia cultural, entregue una serie de etiquetas en base a las características geométricas más importantes del mismo, contribuyendo así al proceso de etiquetado manual de patrones por parte de expertos de la arqueología como una herramienta de apoyo al proceso, sentando en mismo tiempo un precedente para una futura herramienta que lo realice de forma completamente automática.

Objetivos Específicos

1. Obtener un conjunto de patrones etiquetados por un experto a partir del procesamiento del libro *Ornamente Geometric Vasen: Ein Kompendium* [11].
2. Obtener, a partir de los datos iniciales, un conjunto refinado y preciso de datos, apto para ser utilizado como entrada de modelos y algoritmos de aprendizaje supervisado.
3. Encontrar, mediante la experimentación con múltiples técnicas, un modelo de aprendizaje multietiqueta que permita asignar etiquetas probables a los patrones geométricos, basándose para ello en el entrenamiento sobre los datos mencionados.
4. Utilizar el trabajo realizado y la experiencia ganada para establecer, en base a los resultados, una serie de razonamientos que permitan alcanzar un mayor entendimiento del problema del etiquetado de patrones geométricos y la factibilidad de abordarlo desde un enfoque de aprendizaje multietiqueta.
5. Concluir con un análisis de los fenómenos descubiertos, las dificultades encontradas y posibles lineamientos para un trabajo futuro.

Evaluación

Para el desarrollo y evaluación del trabajo se utilizará un conjunto de patrones previamente etiquetados extraídos desde el libro *Ornamente Geometric Vasen: Ein Kompendium* [11],

obra del prof. Dr. Norbert Kunisch, arqueólogo de renombre mundial y ex director del Departamento de Antigüedades del Museo de Ruhr-Universität en Bochum, Alemania. Dicha obra establece una taxonomía para el etiquetado mediante patrones encontrados en las excavaciones más grandes del siglo XX en Grecia, por lo cual se considera un buen referente tanto en cuanto a expertiz del etiquetador como a relevancia de los patrones estudiados.

Los datos serán entonces divididos en subconjuntos de entrenamiento y prueba, y se utilizarán para entrenar distintos algoritmos y modelos de aprendizaje supervisado. De esta manera, la solución ideal será aquella que, tras haber entrenado usando el primer subconjunto, logre una predicción perfecta en el conjunto de prueba, es decir, asigne todas las etiquetas correspondientes - simulando a la perfección el trabajo que habría realizado el etiquetador experto -.

Más formalmente, las técnicas a utilizar se evaluarán mediante distintas métricas de efectividad, siendo las dos principales el *Hamming Score* y el *Recall* promediado sobre las distintas etiquetas. Cabe mencionar que la elección de *Hamming Score* por sobre otras métricas más tradicionales (*accuracy*, *precision*, etc) se debe a la naturaleza misma del problema de clases múltiples y cómo la métrica escogida se adapta mejor, según lo comentado en la Sección 2.

Más en profundidad, *Hamming Score* (cuya formulación matemática se presentó en la Ecuación 1) se define como la proporción de etiquetas correctamente predecidas con respecto al número total de etiquetas (predichas y reales) para una instancia, promediado sobre todas las instancias. Esto resulta particularmente útil para el problema en estudio, en cuanto entregará una noción de qué tan bien se están prediciendo las etiquetas de cada patrón, dando paso a la posibilidad de tener patrones parcialmente bien etiquetados (a diferencia de lo que sucedería, por ejemplo, con *Exact Match Ratio* [5]).

Recall, por otra parte, se define como la proporción de etiquetas correctamente asignadas con respecto al total de etiquetas asociadas, representando así qué tan bien se "cubren" las etiquetas correspondientes sin importar si en el proceso se asignan etiquetas incorrectas (véase Ecuación 2). Esto responde en particular a la hipótesis de que a los expertos les importará, por sobre todo, que las etiquetas correctas les sean efectivamente sugeridas, sin importar si para ello deben descartar una cierta cantidad de sugerencias erróneas (pues detectar estas debería significar un esfuerzo menor).

Un último punto a mencionar sobre la Evaluación guarda relación con la extremadamente baja *label density* presente en los datos, lo cual refleja la alta complejidad del problema a tratar. Este fenómeno, que será detallado más en profundidad en la Sección 4, levanta la necesidad de realizar ciertas refinaciones y supuestos sobre los datos, los cuales llevarán a trabajar sobre un conjunto "podado" de etiquetas donde se considerarán solamente aquellas con al menos t apariciones en los datos de entrenamiento. Dicho valor t se denominará el *umbral de frecuencia*, y su valor tendrá una incidencia directa en los resultados arrojados por cada métrica, en cuanto determinará la escala del problema.

4. Solución Propuesta

La solución propuesta se basa en la concepción de un modelo o algoritmo de aprendizaje multietiqueta que, tomando la imagen de un patrón en escala de grises como entrada, entregue un conjunto de etiquetas Z_i probables para el patrón en cuestión, constituyendo así una herramienta de apoyo al etiquetado manual.

Para lo anterior será necesario definir en primer lugar un conjunto apropiado de datos. Para ello se tomará como restricción de la investigación el uso de los patrones disponibles en el libro *Ornamente Geometric Vasen: Ein Kompendium,* obra magna del etiquetado de patrones geométricos en objetos de herencia cultural, la cual propone una taxonomía a través de 776 patrones asociados a 586 etiquetas distintas disponibles en inglés, alemán, francés y griego.

Dichos datos se utilizarán para entrenar algoritmos y modelos de aprendizaje supervisado desde dos grandes enfoques: el aprendizaje multietiqueta tradicional y el aprendizaje profundo. Para cada enfoque, se buscará experimentar tanto con los métodos del estado del arte como propuestas propias, a fin de tener un amplio espectro de resultados que permita reconocer no solo la técnica más conveniente si no también generar conjeturas respecto al porqué de dicho resultado, dando así lugar a lineamientos para un trabajo futuro.

Un punto importante a mencionar sobre los datos es que, en su estado original, las 586 etiquetas distintas se distribuyen a una tasa de 4.6 etiquetas por patrón, lo cual da como resultado un conjunto de datos con una *label density* [21] de 0.0052, valor extremadamente desfavorable para el aprendizaje de clasificación múltiple [2] y que refleja la presencia de etiquetas con frecuencias muy bajas en los datos. Esto llevado a un extremo tal que, por ejemplo, el 60 % de las etiquetas no aparece más de una vez en todo el conjunto (imposibilitando así su aprendizaje).

Es debido a lo anterior que, antes de aplicar cualquier algoritmo o modelo, será necesario llevar a cabo un tratamiento de datos y tomar algunas hipótesis que permitan simplificar el problema hasta un punto tratable, mas sin perder el valor que la solución pueda representar para el área. Bajo este objetivo, se propone la aplicación simultánea de dos enfoques: una refinación de las etiquetas mediante técnicas de procesamiento de lenguaje natural (particularmente, *lemmatization*) y una "poda" de etiquetas mediante la aplicación de un umbral sobre la frecuencia en entrenamiento. Ambos enfoques se describirán a continuación.

En primer lugar, el uso de técnicas de procesamiento de lenguaje natural para refinar los datos responde a la existencia de etiquetas con características del lenguaje consideradas como innecesarias y que de ser suprimidas podrían reunirse con otras etiquetas, aumentando así la ya mencionada *label density*. Una de ellas es la coexistencia de etiquetas compuestas y sus elementos como etiquetas individuales, que se hace presente, por ejemplo, en las etiquetas *as ornament* y *ornament*, las cuales de reunirse en una sola opción aumentarían su frecuencia al mismo tiempo que disminuirían la cantidad total de etiquetas. Otra dificultad se encuentra en la diferencia de género, problema que se hace particularmente presente para la versión en francés de las etiquetas. Sin embargo, esto último se soluciona tomando la versión en inglés (lenguaje neutro).

Por otro lado, la necesidad de "podar" el problema responde a la infactibilidad de obtener una *label density* suficientemente alta con la sola aplicación de *lemmatization*, en vista de que una gran cantidad de etiquetas seguirán presentando una frecuencia considerablemente baja. En este sentido, la poda corresponde a eliminar de las etiquetas por aprender aquellas que presenten una frecuencia en los datos de entrenamiento menor a t, valor definido como el *umbral de frecuencia,* y que deberá ser determinado de manera que (1) permita alcanzar una *label density* razonable y (2) no simplifique el problema más allá de lo útil, en cuanto valores de umbral demasiado altos podrían dejar muy pocas etiquetas por predecir (cayendo incluso, en el caso más extremo, en una clasificación singular).

Una vez definido el tratamiento de los datos, sigue el detalle de las técnicas a evaluar. Al respecto:

- Para el enfoque de MLC tradicional, se buscará cubrir las dos familias de algoritmos presentados en el Estado del Arte (Sección 2), para lo cual se realizarán experimentos con BR, LP, CC, RakelD, MLkNN, BRkNN y MLTSVM como métodos de aprendizaje multietiqueta, cada uno de los cuales será evaluado con regresión logística, árboles de decisión, máquinas de soporte vectorial y Naïve Bayes como clasificadores base.
- Siguiendo el mismo enfoque, ambas familias de algoritmos serán puestas a prueba tomando como entrada descriptores obtenidos a partir de arquitecturas ResNet18 y ResNet50 entrenadas bajo la tarea de predecir el capítulo del libro al que cada patrón pertenece (5 capítulos). Dicho experimento forma parte de una investigación previa realizada por los profesores guía y su uso en este punto se toma como restricción de la investigación.
- En lo que respecta al enfoque de aprendizaje profundo, se realizarán experimentos tanto con redes *shallow* como redes más avanzadas, a fin de seguir así las recomendaciones aportadas por la literatura (es decir, comenzar con arquitecturas que permitan un entrenamiento y predicción eficaces, e ir avanzando progresivamente hacia arquitecturas más pesadas en función de los resultados).
- Es importante mencionar que, para ambos enfoques, se complementará el conjunto de datos con ejemplos creados sintéticamente mediante múltiples técnicas de aumentado de imágenes, lo cual se espera permita, además de lidiar con la baja cantidad de datos, identificar aquellas técnicas de creación de imágenes sintéticas que mejor se adapten al problema en cuestión.

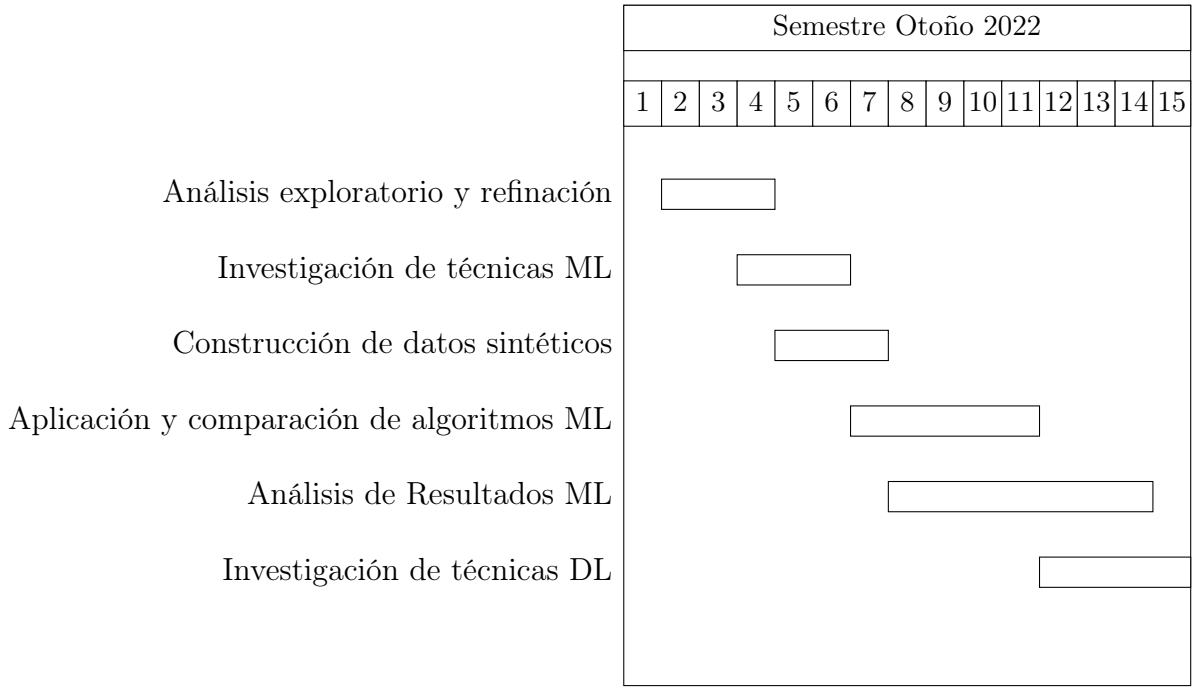
Finalmente, aquel modelo o algoritmo que obtenga los mejores resultados en la tarea de predecir etiquetas para un patrón dado según las métricas definidas en la Sección 3 será aquel considerado como la herramienta de apoyo al etiquetado. Se pretende además que el análisis conjunto de lo obtenido a partir de cada enfoque, así como la experiencia adquirida durante su aplicación, permita dar lugar a distintas conjeturas sobre el porqué de los resultados y, de esta forma, generar lineamientos para un trabajo futuro con miras en la automatización total del proceso.

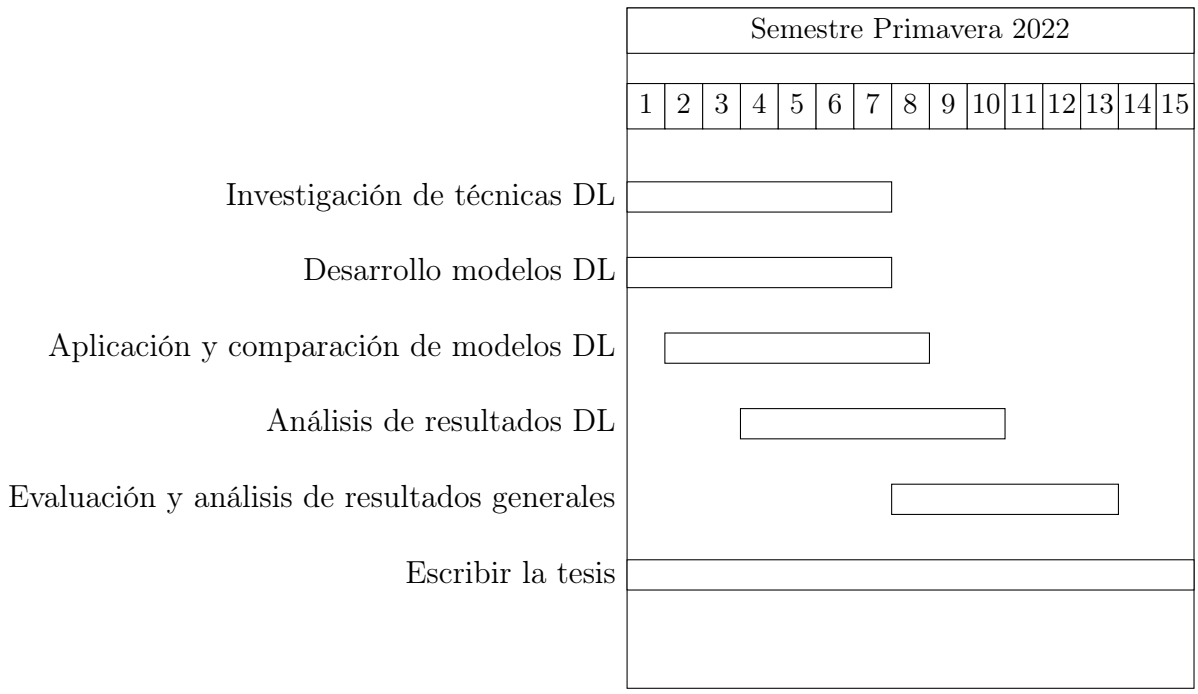
5. Plan de Trabajo

El plan de trabajo contempla las siguientes etapas:

1. Extracción de datos a partir del libro Ornamente Geometrischen Vasen: Ein Kompendium.
2. Análisis exploratorio de datos y refinación mediante técnicas de NLP.
3. Investigación del estado del arte para el problema de etiquetado múltiple mediante un enfoque tradicional.
4. Construcción de conjuntos de datos sintéticos a partir de múltiples transformaciones lineales y no lineales.
5. Selección, aplicación y comparación de algoritmos del enfoque tradicional sobre conjuntos de datos originales y sintéticos.
6. Investigación del estado del arte para el problema de etiquetado múltiple mediante un enfoque de aprendizaje profundo.
7. Construcción, aplicación y comparación de modelos de aprendizaje profundo sobre conjuntos de datos originales y sintéticos.
8. Evaluación y análisis de resultados obtenidos por cada enfoque y en conjunto.

A continuación se presenta una Carta Gantt con las principales etapas divididas entre los semestres Otoño y Primavera 2022. Las etapas previas a la fecha del presente documento corresponden a trabajo ya realizado, el cual se presentará en la sección siguiente.





6. Trabajo Adelantado

A la fecha, distintas etapas del Plan de Trabajo han sido llevadas a cabo. En esta sección se presentarán los principales resultados obtenidos.

La primera tarea abordada guarda relación con la construcción de un conjunto de datos para la investigación, lo cual se logra mediante la extracción de patrones y etiquetas desde el libro *Ornamente Geometrischen Vasen: Ein Kompendium* [11]. Al respecto, es importante mencionar que si bien la extracción de los patrones (imágenes) no fue necesaria puesto que los profesores contaban con dicho material de antemano, sí fue necesaria la extracción de las etiquetas (texto), lo cual se realizó de forma manual.

Una vez construido el conjunto de datos, se procede a realizar un análisis exploratorio del mismo a fin de entender sus principales características y desafíos. Dicho estudio dio lugar a dos grandes observaciones: en primer lugar, el hecho de que las etiquetas se distribuirían de forma similar a una ley *Zipf*, con algunas pocas etiquetas siendo muy frecuentes en los datos y otras muchas con un número de ocurrencias muy bajo (véase Figura 3). En segundo lugar se encontraría la extremadamente baja *label density* presente en los datos, con un valor de 0,005.

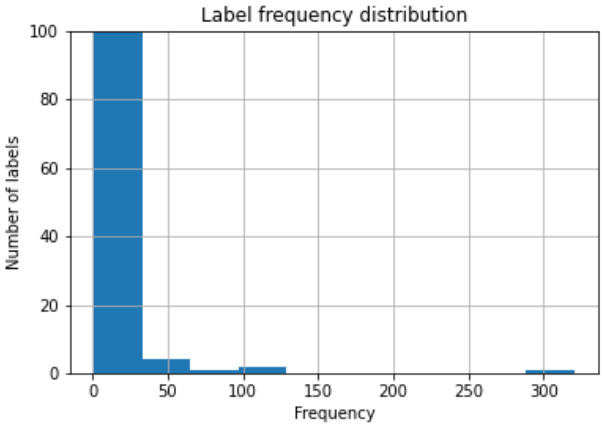


Figura 3: Cantidad de etiquetas v/s frecuencia en el *dataset*.

La presencia de ambas observaciones al mismo tiempo apuntaría a una causa común: la existencia de etiquetas demasiado específicas en los datos. Esto se confirmaría de inmediato al notar que el 82.02 % de las etiquetas no presentan más de 3 apariciones en los datos y, aún más, el 60.02 % de ellas no supera una única aparición.

El fenómeno anterior contribuiría en gran medida a la conformación de la metodología de trabajo, en cuanto revelaría la necesidad de llevar a cabo los experimentos no solamente en un enfoque de método contra método, sino también evaluando el impacto de podar las etiquetas menos frecuentes - reduciendo así la escala del problema -.

Sin embargo, quedaba aún una herramienta por aplicar antes de comenzar a podar etiquetas. Se trataría de *lemmatization*, técnica de NLP que se considerase adecuada para tratar algunas inconsistencias en las etiquetas que daban lugar a duplicados (*as ornament* y *ornament*, por ejemplo). Para ello se aplicaría en primer lugar una limpieza de carácteres innecesarios en los datos (tales como paréntesis y comillas), seguida de una remoción de *stopwords* y finalmente una *lematizacion* mediante *WordNetLemmatizer*, todo lo cual fue posible a través de la librería *Natural Language Toolkit* (NLTK) para Python.

Como resultado del proceso anterior, los datos presentarían ahora un total de 340 etiquetas (en lugar de 586) y una *label density* de 0,015 (tres veces mayor a la original). Se consideró entonces dicho estado como un buen punto de partida para comenzar con un enfoque de podado de etiquetas en base a un umbral frecuencias, el cual se decidió mantener como una variable más en los experimentos a fin de observar cómo el problema se haría más o menos tratable en función de la cantidad de etiquetas a considerar.

Es así como, tras una primera fase de procesamiento de los datos, se daría lugar a los primeros experimentos del enfoque de aprendizaje multietiqueta tradicional. Para ello se opta por utilizar la librería *scikit-multilearn* [18], la cual ofrece un ambiente de Python con todos los métodos multietiqueta del enfoque tradicional.

Los experimentos se llevan entonces a cabo variando el umbral de frecuencias t y evaluando a través del *Hamming Score* la eficacia de cada combinación método-clasificador base listada

en la Sección 4, tomando como entrada descriptores generados con ResNet18 y ResNet50. La Figura 4 presenta el resultado general para ResNet18 (el umbral crece hacia la derecha, lo cual reduce la cantidad de etiquetas y por ende simplifica el problema). La Figura 5, por otra parte, presenta las matrices de confusión obtenidas para cada etiqueta por la mejor combinación algoritmo-clasificador en $t = 25$ sobre descriptores ResNet18 (que corresponde a RakelD-SVC). Los resultados de ResNet50 son similares y no se presentan por economía de espacio.

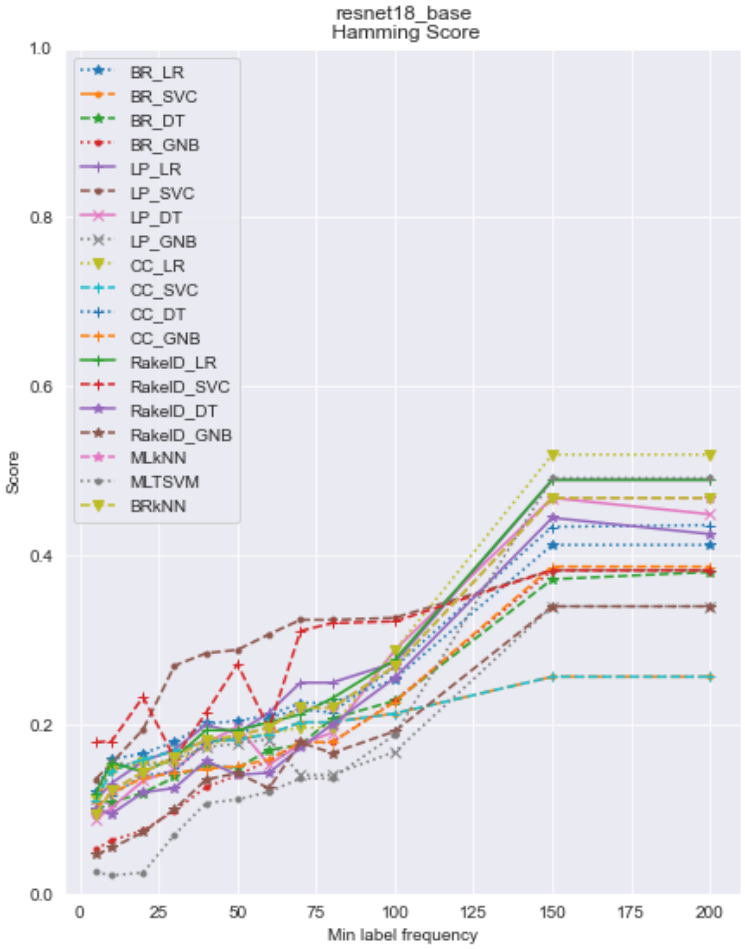


Figura 4: Resultados enfoque tradicional, descriptores de ResNet18 sin *data augmentation*.

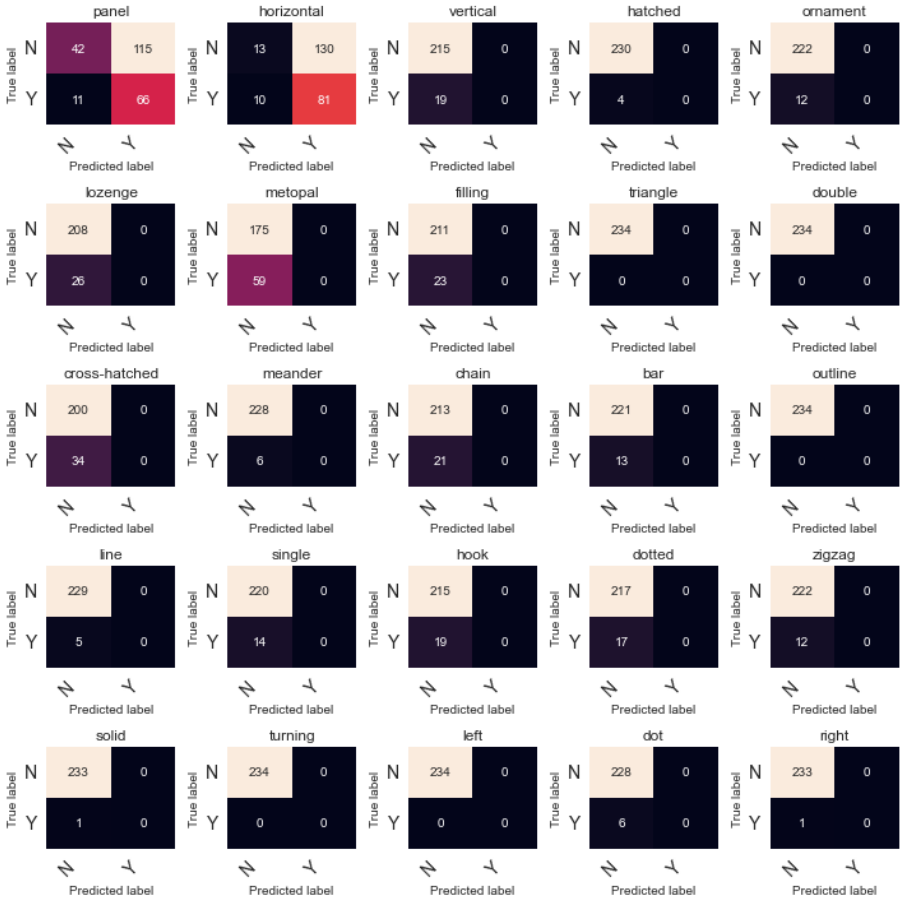


Figura 5: Matriz de confusión 25 etiquetas más frecuentes, RakelD-SVC sobre descriptores de ResNet18 sin *data augmentation*.

Los resultados evidenciarían la capacidad de los métodos por aprender ciertas etiquetas y por ende la existencia de información útil en los descriptores, sin embargo, estarían aún lejos de lo aceptable. Esto debido a que se consideraría que una simplificación "prudente" del problema - es decir, que reduce la complejidad pero manteniendo un desafío cuya solución es útil - es aquella que mantenga alrededor de 25 etiquetas. Dicha condición coincide en este caso con un umbral $t = 25$, valor para el cual todas las combinaciones evaluadas no superan un *Hamming Score* de 0,2418, puntaje demasiado bajo como para dar paso a una herramienta de apoyo. Similar pasa con el *Recall*, puesto que como se observa en las matrices de confusión, solo se estaría prediciendo casos positivos para las 6 etiquetas más frecuentes - respondiendo siempre con la no asignación de la etiqueta en todos los demás casos -.

Lo anterior daría paso a múltiples hipótesis y experimentos, de entre las cuales la más prometedora guardase relación con el desbalance de casos negativos y positivos por cada etiqueta (suma de filas en las matrices de confusión) y cómo este podría estar llevando a los clasificadores base a escoger siempre la clase negativa por estar sobre representada. Esta hipótesis abriría paso para un experimento adicional de BR, en el cual se le entregaría a cada clasificador tantos ejemplos negativos como positivos de cada etiqueta. Dicho experimento daría como resultado matrices de confusión completamente distintas (véase Figura 6) en donde el *Recall* aumentaría pero el *Hamming Score* caería aún más, llegando a alcanzar un

valor de 0,0862.

Sin embargo, el estudio de las matrices de confusión de este último experimento permitiría formular una nueva hipótesis, bajo la cual el problema residiría en que, al buscar entregar a cada clasificador base la misma cantidad de casos negativos y positivos, la cantidad de casos negativos resultante es demasiado baja para lograr representar la enorme variabilidad que estos presentan. Como resultado, las técnicas pierden en parte su capacidad para distinguir el caso negativo, el cual sin embargo sigue estando sobre representado en el conjunto de prueba y es entonces la fuente de un *Hamming Score* tan bajo.

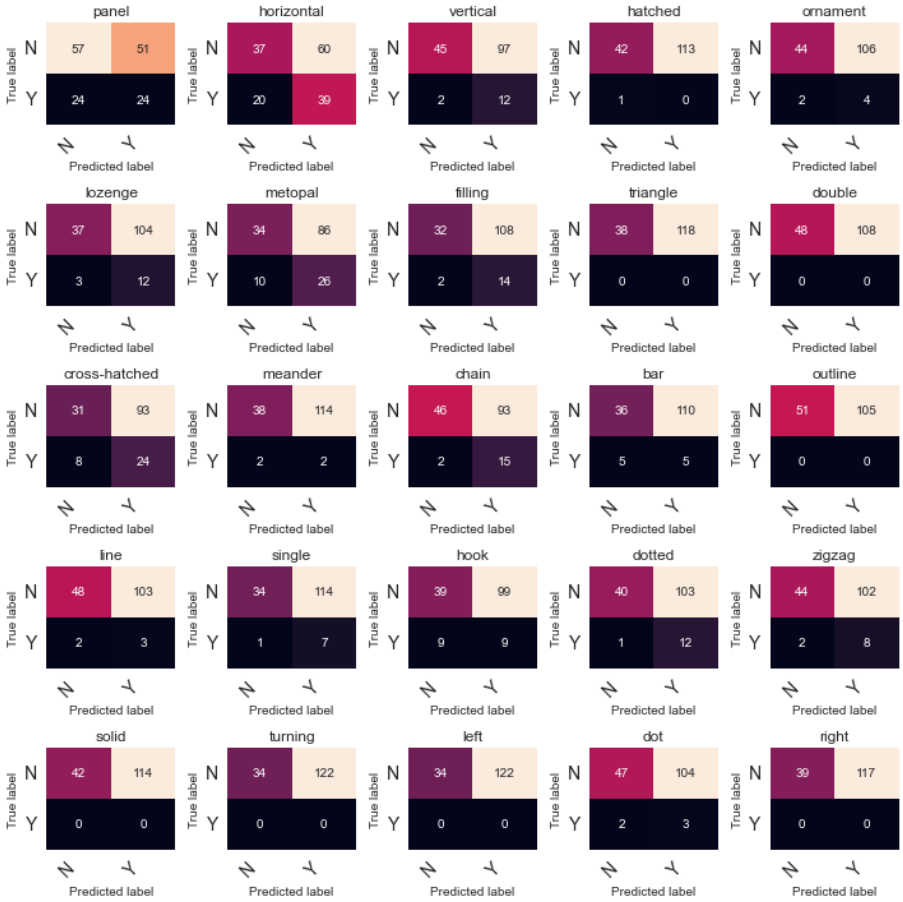


Figura 6: Matriz de confusión 25 etiquetas más frecuentes, BR-LR balanceado sobre descriptores de ResNet18 sin *data augmentation*.

Dicha hipótesis resultaría razonable, en cuanto el problema que describiese era bien conocido y guardaría estrecha relación con la baja *label density* del conjunto de datos. Se optaría entonces por buscar tratar este desbalance mediante la creación de datos sintéticos a través de distintas técnicas de aumentación de imágenes, la mayor parte de ellas implementadas en la librería *imgaug* [10]. Para cada técnica, se reentrenaría la red encargada de generar descriptores durante 200 épocas y se repetirían los experimentos. La Tabla 1 muestra los primeros resultados de dicho proceso para la comparación desbalanceada de métodos.

Al estudiar los resultados obtenidos, llamaba la atención el hecho de no ver mejorías, por

parciales que fuesen, con prácticamente ninguna técnica (siendo la unión de reflexiones con recortes y filtros de difuminado la única excepción). Aún más, a medida que el *accuracy* de la red encargada de generar descriptores aumentaba gracias a los nuevos ejemplos, el *Hamming Score* en general disminuía.

	ResNet18			ResNet50		
D.A. Techniques	Test Accuracy	Hamming Score w/ 25 labels	Best algorithm	Test Accuracy	Hamming Score w/ 25 labels	Best algorithm
'base' (no D.A.)	0.9103	0.2418	LP SVC	0.7564	0.2160	RakelD SVC
'ref'	0.7885	0.1877	LP SVC	0.7500	0.2037	LP SVC
'rot'	0.8397	0.1968	CC SVC	0.8590	0.1719	LP SVC
'blur'	0.7885	0.1906	LP SVC	0.7885	0.1979	RakelD SVC
'crop'	0.8205	0.1788	LP SVC	0.7885	0.1821	BR SVC
'elastic'	0.8141	0.2376	RakelD SVC	0.7756	0.2111	RakelD LR
'rain'	0.8205	0.1828	BR LR	0.8269	0.1828	BR SVC
'random'	0.8269	0.2185	RakelD SVC	0.7885	0.1690	CC SVC
'crop' x2	0.7949	0.1835	BR SVC	0.7949	0.1988	RakelD LR
'elastic' x2	0.8077	0.2376	RakelD SVC	0.7564	0.2142	LP SVC
'random' x2	0.8333	0.1968	CC SVC	0.8397	0.2003	BR LR
'ref', 'rot', 'blur'	0.8077	0.1997	LP SVC	0.8269	0.1779	LP SVC
'ref', 'rot', 'crop'	0.7821	0.1855	RakelD LR	0.7564	0.1864	LP SVC
'ref', 'crop', 'blur'	0.8205	0.1920	RakelD LR	0.7949	**0.2456**	RakelD SVC
'crop' x2, 'elastic' x3 'rain'	0.8333	0.1844	BR LR	0.7885	0.2022	BR LR
'crop' x2, 'elastic' x3 blur'	0.8654	0.2204	BR LR	0.8590	0.1732	RakelD SVC
'ref', 'crop', 'blur', 'rain'	0.8590	0.1963	RakelD SVC	0.7179	0.1673	RakelD SVC

Tabla 1: Resultados tras aplicar *Data Augmentation*, ambas redes, caso *t=25*.

Lo anterior daría paso a una última hipótesis para el enfoque tradicional, y que guardaría relación con la principal desventaja del mismo: una fuerte dependencia con los descriptores, ajenos al proceso de aprendizaje. En particular, se propondría que el principal obstáculo al aprendizaje radicaría en una especialización de las redes generadoras para el problema para el cual están siendo entrenadas (predicción de capítulos en el libro), alejándose así de la capacidad de distinguir una taxonomía más al detalle (como lo son las etiquetas).

A fin de validar dicha hipótesis, se reentrenarían modelos de ResNet18 y ResNet50 sobre el conjunto de datos base y aumentado con las técnicas *blur* (difuminado) y *elastic* (deformación elástica) por separado, tomando salidas del modelo cada 5 épocas y utilizándolas para generar descriptores que posteriormente fueron evaluados tal como se presentó en la Figura 4, prestando especial atención al caso $t = 25$. La Tabla 6 presenta los resultados para los datos aumentados mdiante *blur*.

En los tres casos estudiados, el resultado sería el mismo: mientras más entrenado se encontrase el modelo con el cual se generaban los descriptores, menor era la probabilidad de observar un *Hamming Score* mejor que el resultante de entrenar por 200 épocas. Se observaría además que, a partir de las 90 épocas, dicha métrica no hacia si no decaer y, sin embargo, el *test accuracy* - métrica que habría determinado qué salida tomar en los experimentos de la Tabla 6 - seguía aumentando.

La hipótesis propuesta era entonces validada, pues el hecho de que el *Hamming Score* resultase más alto en épocas tempranas reflejaba que los *embeddings* resultantes de un entrenamiento parcial se adaptaban mejor al problema multietiqueta que aquellos de un entrenamiento prolongado, lo cual evidenciaba el aprendizaje de una representación muy buena para el problema de la red pero demasiado específica para la tarea multietiqueta.

	ResNet18			ResNet50		
Epoch	Test Accuracy	Hamming Score top 25 labels	Best Algorithm	Test Accuracy	Hamming Score top 25 labels	Best Algorithm
Best	0.7885	0.1906	LP SVC	0.7885	0.1979	RakelD SVC
5	0.3974	**0.2368**	RakelD SVC	0.3974	0.1809	RakelD SVC
10	0.6474	0.1752	BR LR	0.6979	0.1838	RakelD SVC
15	0.5833	0.1880	LP SVC	0.5833	0.1826	BR LR
20	0.6603	0.1754	BR LR	0.6603	0.1827	RakelD SVC
25	0.7115	0.1796	RakelD LR	0.7115	0.1767	BR SVC
30	0.7564	**0.2521**	RakelD SVC	0.7564	**0.2093**	RakelD SVC
35	0.6410	0.1995	RakelD SVC	0.6410	0.1992	RakelD SVC
40	0.7179	0.2016	LP SVC	0.7179	**0.2361**	RakelD SVC
45	0.7564	0.2085	RakelD SVC	0.7564	0.1943	RakelD SVC
50	0.7436	**0.2381**	RakelD SVC	0.7414	0.1903	BR LR
55	0.7500	**0.2470**	RakelD SVC	0.7500	0.2098	RakelD SVC
60	0.6795	0.1932	RakelD SVC	0.6795	**0.2225**	RakelD SVC
65	0.6859	0.2061	BR LR	0.6859	0.1694	RakelD SVC
70	0.7885	0.1849	RakelD SVC	0.7885	0.2057	RakelD SVC
75	0.7692	**0.2621**	RakelD SVC	0.7692	0.2019	RakelD SVC

Tabla 2: Resultados de salidas parciales para la técnica *blur*, ambas redes, caso $t = 25$.

Lo anterior sería un buen indicador de que la investigación habría llegado hasta los límites del enfoque tradicional, en cuanto el obstáculo encontrado correspondería a una desventaja inherente al mismo. Ante ello resultaría natural tomar el camino adoptado por la literatura, es decir, dejar a un lado este enfoque para estudiar el problema desde el aprendizaje profundo, abriendo así la posibilidad de incorporar la generación de representaciones en el proceso de aprendizaje.

De esta manera, el trabajo realizado hasta la fecha concluye con el ya presentado estudio del estado del arte para el enfoque de aprendizaje profundo, cerrando así el enfoque tradicional tras haber sentado un conjunto de hipótesis y resultados que servirán de guía y base para los nuevos experimentos, siguiendo así la planificación propuesta en la Sección 5.

Referencias

[1] Biasotti, Silvia, Elia Moscoso Thompson, L Bathe, Stefano Berretti, Andrea Giachetti, T Lejemble, Nicolas Mellado, Konstantinos Moustakas, Iason Manolas, Dimitrios Dimou, Claudio Tortorici, Santiago Velasco-Forero, Naoufel Werghi, Martina Polig, Giusi Sorrentino y Sorin Hermon: *SHREC'18 track: Recognition of geometric patterns over 3D models*. Enero 2018.

[2] Blanco, Alberto, Arantza Casillas, Alicia Pérez y Arantza Díaz de Ilarraza: *Multi-label clinical document classification: Impact of label-density.* Expert Syst. Appl., 138, 2019.

[3] Castillo Luján, Feren: *Tipología y seriación de la cerámica proveniente del cementerio chimú de huaca de la luna, Perú.* Boletín del Museo Chileno de Arte Precolombino, 23:27 – 58, 2018, ISSN 0718-6894. http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-68942018000300027&nrm=iso.

[4] Chen, Wei Jie, Yuan Hai Shao, Chun Na Li y Nai Yang Deng: *MLTSVM: a novel twin support vector machine to multi-label learning.* Pattern Recognition, 52:61–74, 2016.

[5] Dhatri Ganda, Rachana Buch: *A Survey on Multi Label Classification.* Recent Trends in Programming languages, 5(1):19–23, 2018. http://computers.stmjournals.com/index.php?journal=RTPL&page=article&op=view&path%5B%5D=1582.

[6] Ghamrawi, Nadia y Andrew McCallum: *Collective multi-label classification.* En Herzog, Otthein, Hans-Jörg Schek, Norbert Fuhr, Abdur Chowdhury y Wilfried Teiken (editores): *Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management, Bremen, Germany, October 31 - November 5, 2005*, páginas 195–200. ACM, 2005. https://doi.org/10.1145/1099554.1099591.

[7] Godbole, Shantanu y Sunita Sarawagi: *Discriminative Methods for Multi-labeled Classification.* En Dai, Honghua, Ramakrishnan Srikant y Chengqi Zhang (editores): *Advances in Knowledge Discovery and Data Mining, 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May 26-28, 2004, Proceedings*, volumen 3056 de *Lecture Notes in Computer Science*, páginas 22–30. Springer, 2004. https://doi.org/10.1007/978-3-540-24775-3_5.

[8] He, Huihui y Rui Xia: *Joint Binary Neural Network for Multi-label Learning with Applications to Emotion Classification.* En Zhang, Min, Vincent Ng, Dongyan Zhao, Sujian Li y Hongying Zan (editores): *Natural Language Processing and Chinese Computing - 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26-30, 2018, Proceedings, Part I*, volumen 11108 de *Lecture Notes in Computer Science*, páginas 250–259. Springer, 2018. https://doi.org/10.1007/978-3-319-99495-6_21.

[9] Huang, Yi, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür y Elif Ozkirimli: *Balancing Methods for Multi-label Text Classification with Long-Tailed Class Distribution.* CoRR, abs/2109.04712, 2021. https://arxiv.org/abs/2109.04712.

[10] Jung, Alexander B.: *imgaug.* https://github.com/aleju/imgaug, 2018. [Online; accessed 30-Oct-2018].

[11] Kunisch, Norbert: *Ornamente Geometrischer Vasen: Ein Kompendium.*

[12] Lee, Jieun, Ji Hyun Yi y Seungjun Kim: *Cultural Heritage Design Element Labeling System With Gamification.* IEEE Access, 8:127700–127708, 2020.

[13] Liu, Weiwei, Xiaobo Shen, Haobo Wang y Ivor W. Tsang: *The Emerging Trends of Multi-Label Learning.* CoRR, abs/2011.11197, 2020. https://arxiv.org/abs/2011.11197.

[14] NHCN: *Heritage objects: The National Heritage Council of Namibia*, 2005. https://www.nhc-nam.org/heritage-objects.

[15] Read, Jesse, Bernhard Pfahringer, Geoffrey Holmes y Eibe Frank: *Classifier Chains for Multi-label Classification*. En Buntine, Wray L., Marko Grobelnik, Dunja Mladenic y John Shawe-Taylor (editores): *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part II*, volumen 5782 de *Lecture Notes in Computer Science*, páginas 254–269. Springer, 2009. https://doi.org/10.1007/978-3-642-04174-7_17.

[16] Silverman, Helaine, D. Fairchild Ruggles y Logan William S.: *Closing Pandora's Box: Human Rights Conundrums in Cultural Heritage*. Springer Science+Business Media, 2007.

[17] Sipiran, Ivan, Patrick Lazo, Cristian Lopez, Milagritos Jimenez, Nihar Bagewadi, Benjamin Bustos, Hieu Dao, Shankar Gangisetty, Martin Hanik, Ngoc-Phuong Ho-Thi, Mike Holenderski, Dmitri Jarnikov, Arniel Labrada, Stefan Lengauer, Roxane Licandro, Dinh-Huan Nguyen, Thang-Long Nguyen-Ho, Luis A. Pérez Rey, Bang-Dang Pham, Reinhold Preiner, Tobias Schreck, Quoc-Huy Trinh, Loek Tonnaer, Christoph von Tycowicz y The-Anh Vu-Le: *SHREC 2021: Retrieval of cultural heritage objects*. Comput. Graph., 100:1–20, 2021. https://doi.org/10.1016/j.cag.2021.07.010.

[18] Szymański, P. y T. Kajdanowicz: *A scikit-based Python environment for performing multi-label classification*. ArXiv e-prints, Febrero 2017.

[19] Tsoumakas, Grigorios y Ioannis Katakis: *Multi-Label Classification: An Overview*. Int. J. Data Warehous. Min., 3(3):1–13, 2007. https://doi.org/10.4018/jdwm.2007070101.

[20] Tsoumakas, Grigorios, Ioannis Katakis y Ioannis Vlahavas: *I.: A Review of Multi-Label Classification Methods*. En *In: Proceedings of the 2nd ADBIS Workshop on Data Mining and Knowledge Discovery (ADMKD)*, páginas 99–109, 2006.

[21] Tsoumakas, Grigorios, Ioannis Katakis y Ioannis P. Vlahavas: *Mining Multi-label Data*. En *Data Mining and Knowledge Discovery Handbook*, páginas 667–685. Springer, 2010.

[22] Tsoumakas, Grigorios y Ioannis P. Vlahavas: *Random* k *-Labelsets: An Ensemble Method for Multilabel Classification*. En Kok, Joost N., Jacek Koronacki, Ramón López de Mántaras, Stan Matwin, Dunja Mladenic y Andrzej Skowron (editores): *Machine Learning: ECML 2007, 18th European Conference on Machine Learning, Warsaw, Poland, September 17-21, 2007, Proceedings*, volumen 4701 de *Lecture Notes in Computer Science*, páginas 406–417. Springer, 2007. https://doi.org/10.1007/978-3-540-74958-5_38.

[23] Xioufis, Eleftherios Spyromitros, Grigorios Tsoumakas y Ioannis P. Vlahavas: *An Empirical Study of Lazy Multilabel Classification Algorithms*. En Darzentas, John, George A. Vouros, Spyros Vosinakis y Argyris Arnellos (editores): *Artificial Intelligence: Theories, Models and Applications, 5th Hellenic Conference on AI, SETN 2008, Syros, Greece, October 2-4, 2008. Proceedings*, volumen 5138 de *Lecture Notes in Computer Science*, páginas 401–406. Springer, 2008. https://doi.org/10.1007/978-3-540-87881-0_40.

[24] Yeh, Chih Kuan, Wei Chieh Wu, Wei Jen Ko y Yu Chiang Frank Wang: *Learning Deep Latent Spaces for Multi-Label Classification*. Julio 2017.

[25] Zhang, Min-Ling, Yu-Kun Li, Xu-Ying Liu y Xin Geng: *Binary relevance for multi-label learning: an overview*. Frontiers Comput. Sci., 12(2):191–202, 2018. https://doi.org/10.1007/s11704-017-7031-7.

[26] Zhang, Min-Ling y Zhi-Hua Zhou: *Multi-Label Neural Networks with Applications to Functional Genomics and Text Categorization*. IEEE Trans. Knowl. Data Eng., 18(10):1338–1351, 2006. https://doi.org/10.1109/TKDE.2006.162.

[27] Zhang, Min-Ling y Zhi-Hua Zhou: *A Review on Multi-Label Learning Algorithms*. IEEE Trans. Knowl. Data Eng., 26(8):1819–1837, 2014. https://doi.org/10.1109/TKDE.2013.39.