Universidad Nacional de Luján

Clasificación de flujos de datos continuos y multi etiquetados

Tesina de grado presentada para optar al título de
Licenciado en Sistemas de Información

Juan Cruz Cardona

Director: Santiago Banchero

2021

CLASIFICACIÓN DE FLUJOS DE DATOS CONTINUOS Y MULTI ETIQUETADOS

La clasificación multi-etiquetas es un paradigma de aprendizaje supervisado que generaliza las técnicas clásicas de clasificación para abordar problemas en donde cada instancia de una colección se encuentra asociada a múltiples etiquetas. La mayor parte de los trabajos de investigación han sido realizados en contextos de aprendizaje por *batch*. Los ambientes de flujo continuo de datos (o *streaming*) presentan nuevos desafíos a esta área debido a las limitaciones de tiempo de respuesta y almacenamiento que acarrean. A esto se agrega la naturaleza evolutiva de este tipo de escenarios, que obligan a los algoritmos a adaptarse a cambios de concepto. En la presente investigación se aplican algoritmos de clasificación multi-etiquetas a colecciones estructuradas y no estructuradas. Los experimentos se llevarán a cabo en ambientes simulados de *streaming* de datos para conocer el impacto que produce este contexto sobre los resultados de la clasificación y acoplar el modelo a escenarios del mundo real. A su vez, se partirá de estas colecciones de datos para generar instancias sintéticas y así producir flujos potencialmente infinitos. Por último, se abordarán estrategias de ensambles de algoritmos en búsqueda de una mejora en la calidad de la tarea de predicción de objetos no observados por el modelo. De esta manera, se proveerá a la comunidad de nuevos estudios experimentales sobre algoritmos y colecciones ya conocidos del área de clasificación multi-etiquetas, de manera tal de extender el conocimiento sobre su rendimiento bajo escenarios evolutivos y de naturaleza variable.

Palabras claves: clasificación, multi-etiquetas, *streaming*, algoritmos, flujos.

Índice general

1. INTRODUCCIÓN

1.1. Fundamentos

En los últimos años ha habido un aumento considerable de datos de diversa índole y generados por fuentes heterogéneas. Según los autores Gantz y Reinsel, el volumen total de datos creados y replicados en el mundo durante el año 2011 supera los 1.8 ZB (zettabytes) y se ha estimado que duplica cada dos años [4]. Los avances en el área de tecnología de la información (IT) han contribuido a una continua producción de datos y expansión del campo digital, tal es el caso para la red social *Facebook*, la cual recibe cada hora un flujo de 10 millones de fotos que publican sus usuarios [7]. A estas grandes colecciones de datos se las conoce como *big data* y acarrean nuevas oportunidades y desafíos al campo de las ciencias de la computación. En cuestiones económicas, un análisis a gran escala en búsqueda de tendencias en el comportamiento de los usuarios o clientes de un sistema puede dar una ventaja competitiva en el mercado y, en adición, proveer de un servicio valioso a la comunidad. Potencialmente, la *big data* puede ser una fuente que proporcione a la comunidad de conocimiento nuevo sobre el mundo en el que habita, o como ha mencionado Fayyad, Piatetsky-Shapiro y Smyth en su escrito sobre el descubrimiento de conocimiento, "Los datos que percibimos de nuestro ambiente son la evidencia básica que usamos para construir teorías y modelos sobre el universo en el que vivimos"[1].

Sin embargo, volúmenes masivos de datos tornan obsoletos los tradicionales métodos manuales de análisis de datos y surge la necesidad de desarrollar técnicas automatizadas para extraer patrones en los datos y obtener conocimiento. Con este fin, se han desarrollado técnicas en las áreas de minería de datos y aprendizaje de máquinas que abordan estas colecciones en búsqueda de conocimiento válido y útil. Dichas técnicas se han enfocado en el aprendizaje por *batch* [3], lo que significa que el algoritmo dispone de la colección completa, almacenada en disco, y con la cual genera un modelo a partir de una o múltiples iteraciones sobre todos los datos. No obstante, el aprendizaje por *batch* trae aparejada una dificultad en su misma definición: requiere de todos los datos de la colección presentes y accesibles en todo momento, lo cual no siempre es posible. Además se suma una limitante que es clave en el contexto actual de alta disponibilidad de datos: hoy en día una buena parte de los datos generados proviene de flujos continuos o '*streamings*' de datos [1]. Estos flujos son potencialmente ilimitados, arriban de a una instancia por vez, y son analizados con restricciones altas de tiempo de procesamiento y de memoria. Tal es el caso para aplicaciones de sensores, monitoreo de redes y administración de tráfico, flujo de clics de un usuario en la web, redes sociales, entre otros. Los algoritmos de aprendizaje que actúen en este entorno dinámico deben contar con mecanismos que permitan manejar cambios en la naturaleza o distribución de los datos, tanto para incorporar datos nuevos, como para descartar los datos antiguos. Por estas razones, se torna necesario que las aplicaciones basadas en clasificación en tiempo real adapten sus operaciones de entrenamiento y predicción para lograr mejores resultados [12].

Dentro del área de minería de datos, una de las principales tareas es la de clasificación, la cual consiste en entrenar un modelo que sea capaz de asignar una única etiqueta a una instancia desconocida. No obstante, existen problemas de clasificación en donde múltiples etiquetas son necesarias para caracterizar una instancia. Por ejemplo, una noticia de diario

[1] "Data we capture about our environment are the basic evidence we use to build theories and models of the universe we live in" [2, p. 2]. Traducción propia.

referida al accidente aéreo que sufrió el plantel de fútbol del club Chapecoense puede ser clasificado en la categoría de "Fútbol" tanto como en la de "Tragedias". Del mismo modo, un video documental sobre la vida de Borges puede anotarse como "Biografía", "Literatura" o incluso "Buenos Aires" si se mostraran imágenes de la ciudad. Este tipo de problemas es llamado Clasificación multi-etiquetas (MLL) [2] y representa un nuevo paradigma de aprendizaje automático, con sus propios retos por afrontar y que aún no ha sido suficientemente explorado en proyectos de investigación.

Una clasificación multi-etiqueta permite conocer el grado de correlación entre una instancia de la colección y una o más etiquetas. Esta cualidad significa un mayor poder de generalización con respecto a la clasificación tradicional de única etiqueta, ya que puede abarcar esos mismos problemas y otros de mayor número de etiquetas. Además existen algoritmos que aprovechan la correlación entre etiquetas para mejorar la eficiencia de la clasificación y la calidad de la predicción.

El campo de MLL se ha desarrollado considerablemente en los últimos años pero hasta el momento muchos de estos trabajos se han llevado a cabo en ambientes estáticos de aprendizaje por *batch* [10]. Por lo que se hace necesario encarar nuevos proyectos que aborden clasificaciones MLL en contextos de *streaming* de datos. El desafío entonces consiste en crear clasificadores que sean capaces de manejar un inmenso número de instancias y adaptarse al cambio, a la vez que estar preparados para hacer tareas de predicción en cualquier momento, y todo esto en un contexto de altas restricciones de tiempo de respuesta y memoria.

1.2. Descripción del tema de estudio

1.2.1. Clasificación Multi-etiquetas

Tradicionalmente, el aprendizaje supervisado ha consistido en asociar una instancia o ejemplo a una única etiqueta. Dicho ejemplo es una representación de un objeto del mundo real, y por lo tanto, consta de características o *features* particulares. La etiqueta corresponde a un significado semántico o concepto que lo caracteriza. La tarea de clasificación entonces, reside en aprender una función que permita enlazar ejemplos no observados con una etiqueta. Es preciso notar aquí que dicha definición encubre la restricción de que cada instancia pertenece a una única etiqueta, o dicho de otra manera, cada objeto del mundo real se asocia a un único concepto y ningún otro. Sin embargo, existen problemas de clasificación donde más de una etiqueta puede ser asignada a un ejemplo. La anterior presunción no se amolda a problemas complejos donde un objeto pueda tener más de un significado simultáneamente.

Tareas de este tipo pueden surgir en áreas como las de categorización de texto, recuperación de información musical, clasificación semántica de escenas, anotación automática de videos o clasificación de genes y funciones proteicas. A modo de ejemplo, en el campo mencionado de clasificación semántica de escenas, la foto de un paisaje que ilustra una montaña y una playa puede asociarse a las categorías de 'playa' y 'montaña', simultáneamente [5]; en bioinformática, cada gen puede ser asociado a clases según su función, tales como 'metabolismo', 'transcripción' o 'síntesis proteica' [14]; por último, en recuperación de información musical una pieza sinfónica puede tener *tags* como 'Mozart', 'piano' o 'clásica'.

Este nuevo paradigma es llamado 'Clasificación multi-etiquetas' y ataca problemas con las siguientes características [5]:

[2] Siglas provenientes de su abreviación en inglés, Multi-label learning

- El conjunto de etiquetas es previamente definido y tiene un significado interpretable por un humano.
- El número de etiquetas es limitado y no mayor que el número de atributos.
- En caso que el número de atributos sea grande, se debe poder aplicar estrategias de reducción de atributos.
- El número de ejemplos puede ser grande.
- Las etiquetas pueden estar correlacionadas. Esto significa que se pueden aplicar técnicas que exploten estas relaciones con el objetivo de reducir los tiempo de procesamiento de los algoritmos.
- La distribución de los datos puede estar desbalanceada, es decir, que una etiqueta puede tener un mayor número de ejemplos que otras.

Asimismo, surge un desafío a superar: el conjunto de etiquetas posible crece exponencialmente ante cada nueva adición de una etiqueta. Por ejemplo, si se tuvieran 20 etiquetas, la cantidad posible de conjuntos de etiquetas distintos excedería el millón (2^{20}). Esto implica un tamaño exorbitante del espacio de salida y, en consecuencia, costos computacionales altos. En ese sentido, se ha buscado desarrollar algoritmos que aprovechan las correlaciones o dependencias entre etiquetas. Por ejemplo, la probabilidad de que una noticia que contiene los términos 'pelota' y 'gol' sea anotada con la etiqueta 'fútbol' sería mayor que si se etiquetara con la etiqueta 'música'. Zhang y Zhang clasifican estos algoritmos en tres grupos según la estrategia de correlación aplicada [14]:

Estrategia de primer orden La tarea de MLL es dividida en 'q' tareas de clasificación binarias, siendo 'q' el número de etiquetas de la colección.

Estrategia de segundo orden La tarea de MLL se basa en la generación de relaciones de a pares de etiquetas ya sea por *rankings* entre clases relevantes y no relevantes o por interacción entre pares de etiquetas.

Estrategia de alto orden La tarea de MLL considera relaciones de alto orden entre etiquetas.

Las estrategias de primer orden son conceptualmente simples y eficientes pero logran resultados de menor calidad ya que no consideran correlaciones. Las estrategias de segundo orden tienen un mayor poder de generalización pero no todos los problemas de MLL pueden ser abarcados. El último grupo, por su parte, modela las correlaciones más potentes pero conlleva un costo computacional alto.

En el último tiempo, muchos son los algoritmos que han sido desarrollados para atacar el problema de la clasificación MLL. La comunidad de investigación ha aceptado la taxonomía definida por Tsoumakas y Katakis, para estudiar y clasificar los distintos algoritmos de la literatura [5]. La misma propone dos grandes grupos:

Métodos de transformación del problema Este tipo de algoritmos transforman el problema de clasificación MLL en un problema de clasificación tradicional. Ejemplos típicos de este grupo son Binary Relevance (BR) [13] y Classifier Chains (CC) [10]. Ambos convierten la tarea en una de clasificación binaria.

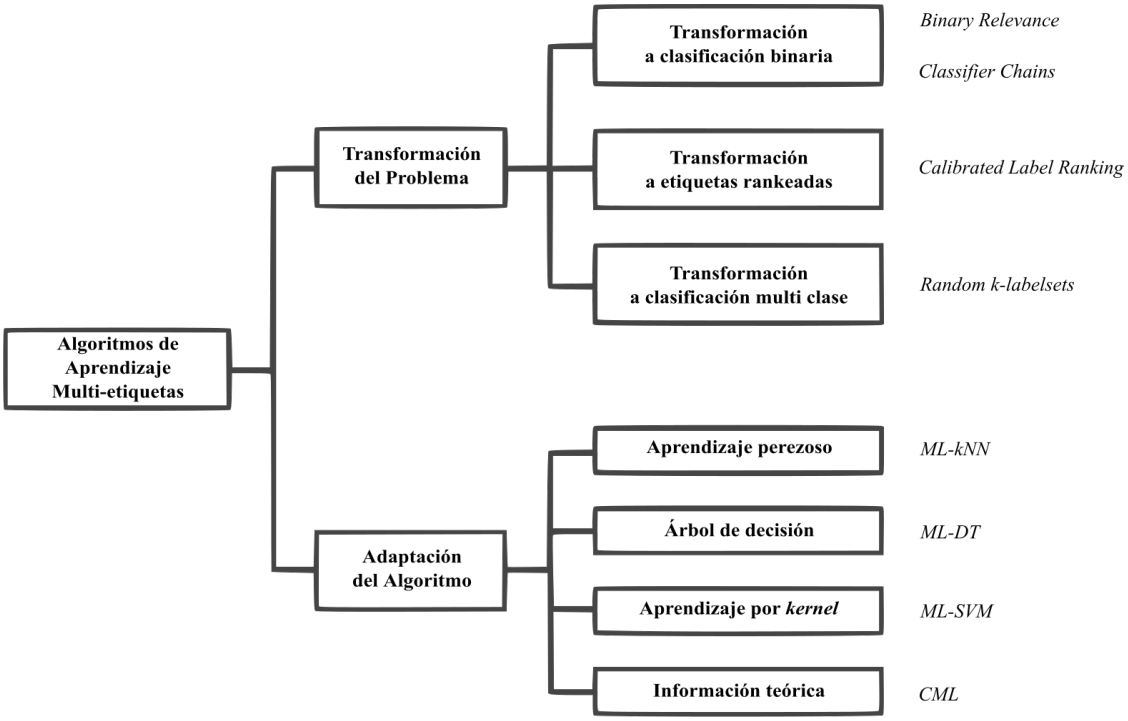

Fig. 1.1: Categorización de los algoritmos de MLL más representativos

Métodos de adaptación del algoritmo Son algoritmos que toman métodos tradicionales de aprendizaje, como árboles de decisión o *naive bayes*, y los adaptan a la tarea de clasificación MLL.

A modo ilustrativo, la figura 1.1 es un diagrama de la taxonomía de algoritmos confeccionado por Zhang y Zhou [15]

1.2.2. Flujos continuos de datos

Hoy en día, los datos pueden ser generados por elementos de continuo monitoreo del medio, tales como sensores o archivos de registro. La clasificación de *streaming* de datos se enfoca en problemas de este tipo, en donde objetos del mundo real son analizados en tiempo real. En este tipo de escenarios, los datos deben cumplir con las siguientes características [3]:

- Los datos están disponibles a través de flujos continuos e ilimitados en el tiempo.
- Las regularidades subyacentes de los datos no son estacionarias sino que pueden evolucionar.
- La data ya no es considerada independiente e idénticamente distribuida.
- La data está situada tanto en el espacio como en el tiempo.

Para hacer frente a estas características de los datos, los sistemas o algoritmos deben contar con los siguientes requerimientos [6]:

- El tiempo para procesar cada registro debe ser constante y pequeño.

- Debe usar un tamaño fijo de memoria principal y no dependiente de la cantidad de registros ya procesados.
- El modelo es generado a partir de una única pasada sobre los datos.
- Debe contar con un modelo listo para realizar predicciones en cualquier momento.
- Idealmente deberá producir un modelo equivalente o casi idéntico al que hubiera sido producido en un ambiente de *batch*.
- Debe mantenerse actualizado ante evoluciones o derivas de concepto en los datos.

Los algoritmos que cumplen con estas cualidades son llamados algoritmos de aprendizaje adaptativo y, aplicados a tareas de clasificación como el de multi-etiquetas, pueden aprovecharse para entrenar un mayor número de objetos y predecir en cualquier instante.

1.3. Motivación

Ante la necesidad de hacer frente a un contexto global de generación masiva de datos y a ritmo acelerado, se hace preciso fortalecer las técnicas de aprendizaje automático actualmente presentes en el campo. En este escenario ya no es posible contar con todos los datos almacenados físicamente y la idea de generar un modelo completo para luego evaluarlo en una fase posterior debe ser reemplazada por una en donde el modelo esté siempre listo para realizar predicciones y al mismo tiempo ser capaz de re-entrenarse y recalcular las métricas de evaluación ante cada nueva instancia abordada. Todo esto en un contexto cambiante, de alta disponibilidad y de limitación en el espacio de almacenamiento. Si bien existen métodos de clasificación para flujos continuos que han dado resultados satisfactorios, aún es un campo conveniente de ser abordado. Asímismo, reproducir los experimentos realizados y fortalecer las técnicas y herramientas actuales puede ser beneficioso para lograr estudios precisos y pormenorizados que sean valiosos tanto para la comunidad científica en sí misma, como también para la sociedad en general.

Por otro lado, si bien existen en el mundo real infinidad de datos multi etiquetados aún no es posible hallar colecciones disponibles al público que cuenten con todas las características de un flujo continuo de datos. Uno de los enfoques abordados es convertir las colecciones existentes en flujos tales que arriben en conjuntos predefinidos y a lo largo del tiempo. De esta manera los algoritmos pueden ser utilizados para realizar clasificaciones en un ambiente similar al de un escenario de *streaming*. Sin embargo, estas colecciones tienen un número limitado de instancias y por lo tanto no cumplen con la condición de ser teóricamente infinitos. Es entonces aquí donde surgen las técnicas de generación sintética de instancias, que buscan reproducir la distribución subyacente de los datos para simular colecciones de datos del mundo real. La contracara de este enfoque es que, si bien existen técnicas y herramientas para generar datos etiquetados, buena parte de ellos son solo aplicables para instancias de una única etiqueta y los que logran generar datos multi etiquetados no han sido lo suficientemente explorados en el área. Hasta el momento, los generadores de instancias multi etiquetadas son capaces de generar datos cercanos a los de colecciones del mundo real [9] y brindan la posibilidad de realizar estudios relativamente certeros de algoritmos de clasificación [11]. No obstante, debe notarse también que si bien se han obtenido colecciones sintéticas en sí mismas aún no han logrado generar instancias para una colección en concreto, respetando sus cualidades particulares y que las distinguen de otras, tales como la co-ocurrencia de etiquetas, la densidad y cardinalidad de las etiquetas y la relación entre las etiquetas y sus *features*, por mencionar algunas. De

lograr esta aproximación se podrán realizar estudios sobre el impacto de los algoritmos sobre flujos de datos de naturaleza distintiva, o en otras palabras, entender en qué medida un algoritmo es más apropiado que otro para un conjunto de datos en un determinado contexto.

Esta última idea mencionada, es decir, el ser capaz de hallar las fortalezas y debilidades de un algoritmo de MLL en un contexto determinado es clave para evaluar la clasificación y entender los resultados obtenidos. Estudios como el de Sousa y Gama [12] o el de Read y col. [11] se han topado con que algoritmos de menor complejidad pueden ser competitivos o incluso superar las métricas de otros algoritmos más complejos. Esta variabilidad en los resultados no solo contrae la necesidad antes mencionada de realizar más estudios al respecto, sino que también abre las puertas a incursionar en soluciones de ensambles de algoritmos. Estos ensambles han dado probada muestra de potenciarse ante la diversidad de resultados obtenidos por sus estimadores base [8], ya que son capaces de disminuir el error total mediante estrategias combinativas. De cualquier manera, las estrategias de ensamble existentes para flujos continuos no han recibido la misma atención que aquellas aplicadas sobre ambientes de *batch* y queda mucho camino por recorrer.

1.4. Objetivos

A partir del marco planteado, el presente trabajo tiene por objetivo principal realizar estudios sobre el impacto de distintos algoritmos de clasificación sobre datos multietiquetados en ambientes de *streaming* provenientes de distintas fuentes de origen y de distinta naturaleza, en particular se seleccionan colecciones de datos que son puntos de referencia en la literatura y que poseen características distintivas entre sí, tales como el número de etiquetas, el número de *features*, o la cantidad de instancias. A su vez, es necesario convertir estos datos a flujos continuos y a este fin se generan instancias sintéticas que sean fieles a estas características mencionadas, aplicando técnicas existentes pero también extendiendolas para detectar co-ocurrencias entre etiquetas. De esta manera, se buscan obtener representaciones óptimas de las colecciones. Finalmente, se llevarán a cabo clasificaciones con algoritmos clásicos y con soluciones de ensambles, en búsqueda de maximizar los valores de las métricas de evaluación en cada escenario. Adicionalmente, se diseñan distintas configuraciones de ensambles, variando los estimadores base y probando distintas implementaciones. Con esto último en mente, se desarrolla una versión del algoritmo de mayoría de voto para el lenguaje Python y se compara su rendimiento contra el de las implementaciones de Java .

describir qué es este algoritmo de mayoría de voto

Con esto en mente, se listan a continuación los objetivos particulares del trabajo:

- Obtener colecciones de datos que cumplan con las propiedades requeridas para considerarse un flujo continuo. Las caracteristicas deben variar entre colecciones. Cada colección debe tener las instancias propias del juego de datos e instancias sintéticas potencialmente ilimitadas.
- Generar flujos continuos de datos a partir de la colección proporcionada replicando su número de etiquetas, features e instancias, su cardinalidad y densidad de etiquetas y la co-ocurrencia entre dos etiquetas.

referencia bibliografica

- Ejecutar algoritmos de clasificación de MLL, para obtener modelos y realizar evaluaciones sobre los resultados. Todo esto sobre distintos escenarios de flujos continuos.
- Proponer una solución de ensambles a partir de la combinación de algoritmos seleccionados de la literatura.

no estan los experimentos todavía, este punto podría cambiar

1.5. Aportes

El presente trabajo de investigación aborda el campo de aprendizaje por multi-etiquetas a partir de la experimentación y evaluación de técnicas y algoritmos de la literatura sobre colecciones de naturaleza cambiante. MLL es un paradigma emergente de aprendizaje supervisado cuyas características implícitas abren paso a nuevos desafíos que derivan del crecimiento exponencial de etiquetas y sus combinaciones, y del costo computacional de entrenar y consultar el modelo. También suelen presentarse otras propiedades como la alta dimensionalidad, data evolutiva y desbalanceada o dependencia entre etiquetas, las cuales implican una re-significación de las técnicas y métodos tradicionales del área de minería de datos.

El paradigma de MLL ha dado muestras de su eficiencia en términos de tiempos de configuración y ejecución de las tareas, bajo diversos campos de aplicación tales como los de categorización de texto, diagnósticos médicos, minería de redes sociales o análisis de datos químicos, y se mantiene en constante expansión hacia nuevos dominios de aplicación. Asimismo, su continua integración a problemas de diversa naturaleza ha contribuido a alimentar esta tendencia.

La tarea de clasificación de *streaming* de datos se enfoca en problemas donde objetos del mundo real son generados y procesados en tiempo real. Datos de este tipo, y que además poseen múltiples etiquetas, son frecuentes en escenarios del mundo real tal como sitios de publicación de imágenes, correos electrónicos o portales de noticias. Abordar este tipo de problemas implica que los algoritmos sean capaces de identificar cambios de concepto en los datos y adaptarse al nuevo contexto. De lograr esto, se podrá generar modelos más sólidos, ya que se cuenta con un mayor número de objetos, y que se encontrarán aptos para predecir en cualquier momento. Surge entonces el reto de crear clasificadores que actúen en ambientes de altas restricciones computacionales y sean capaces de manejar un inmenso número de instancias, lidiar con evoluciones en los datos y estar listos para resolver tareas de predicción en tiempo real.

A diferencia de otros trabajos de investigación recientes, este proyecto lleva a cabo estudios experimentales sobre el tema de clasificaciones multi-etiquetas, para hallar las fortalezas y debilidades de distintos algoritmos de aprendizaje sobre distintos tipos de colecciones, con miras a aportar de un mayor conocimiento empírico sobre el tema a la comunidad científica especializada en tareas de clasificación de flujos de datos multi-etiquetados. El presente trabajo espera contribuir al estudio de métodos y técnicas asentadas, pero también examinar algoritmos exitosos del campo de aprendizaje por *batch*, particularmente los de ensambles de estimadores, a fin de extender su funcionalidad a ambientes de flujos continuos, analizar su desempeño y determinar en qué medida son aptos o no para este tipo de ambientes.

1.6. Organización del Trabajo

Describir las distintas secciones del trabajo

2. PRELIMINARES

En esta sección se presenta el marco teórico de este trabajo, dando un panorama general de cada una de las disciplinas abordadas e introduciendo los conceptos básicos y fundamentales para entender el proyecto. Se comienza con la definición de la taxonomía del campo de estudio, luego describir las siguientes secciones abordadas

2.1. Taxonomía del Campos de Estudio

En pocas palabras, el presente trabajo de investigación se enmarca en las áreas de *big data* y minería de datos, con aplicación en escenarios de *streaming* o flujos continuos de datos y abordando clasificaciones multi-etiquetas. También se aprovechan técnicas del área de procesamiento de lenguaje natural para tratar corpus de texto libre y extraer *features* o características representativas de los datos.

La figura 2.1 es un esquema que ilustra la taxonomía del campo de estudio y la interrelación entre las áreas de investigación involucradas.

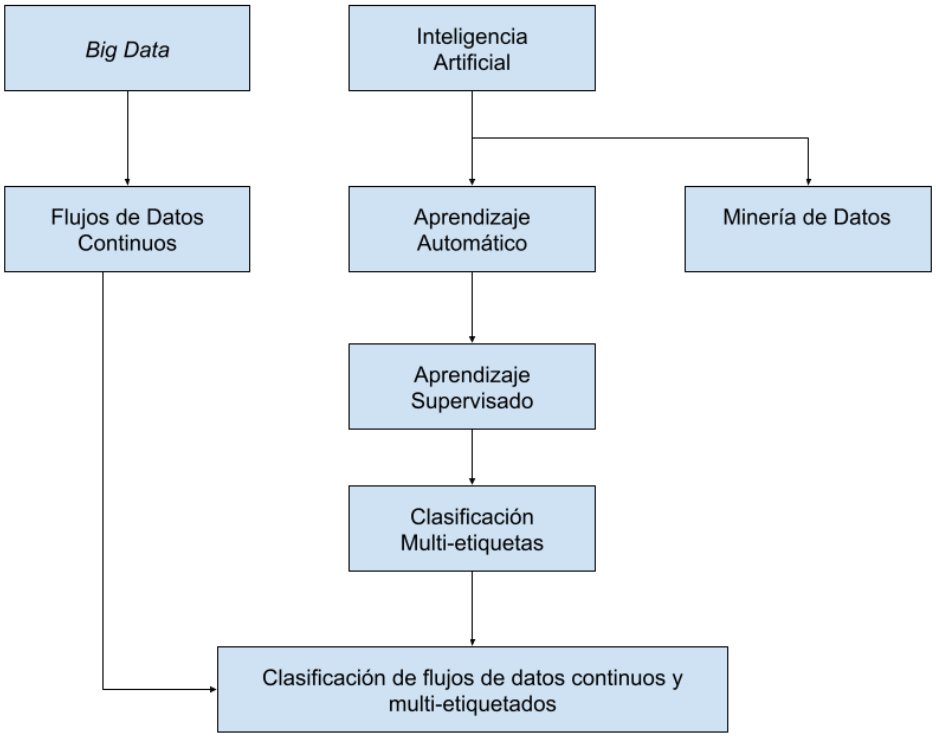

Fig. 2.1: Taxonomía del campo de estudio

MLL Clasificación multi-etiquetas. 2

BIBLIOGRAFíA

[1] Albert Bifet y Gianmarco De Francisci Morales. «Big Data Stream Learning with SAMOA». En: *2014 IEEE International Conference on Data Mining Workshop*. 2014.

[2] Usama M Fayyad, Gregory Piatetsky-Shapiro y Padhraic Smyth. *Advances in Knowledge Discovery and Data Mining*. Ed. por Fayyad, Usama M. and Piatetsky-Shapiro, Gregory and Smyth, Padhraic and Uthurusamy, Ramasamy. Section: From data mining to knowledge discovery: an overview. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996. 1–34.

[3] João Gama. *Knowledge Discovery from Data Streams*. 2010.

[4] J Gantz y D Reinsel. «Extracting value from chaos». En: *IDC IView* (2011), págs. 1-12.

[5] Eva Gibaja y Sebastian Ventura. «A Tutorial on Multi-Label Learning». En: *ACM Computing Surveys* 47 (2015).

[6] Geoff Hulten, Laurie Spencer y Pedro Domingos. «Mining Time-changing Data Streams». En: *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '01. San Francisco, California: ACM, 2001, págs. 97-106.

[7] V Mayer-Schonberger y K Cukier. *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. An Eamon Dolan book. Houghton Mifflin Harcourt, 2013.

[8] Robi Polikar. «Polikar, R.: Ensemble based systems in decision making. IEEE Circuit Syst. Mag. 6, 21-45». En: *Circuits and Systems Magazine, IEEE* 6 (6 de oct. de 2006), págs. 21-45. DOI: 10.1109/MCAS.2006.1688199.

[9] Jesse Read, Bernhard Pfahringer y Geo Holmes. «Generating Synthetic Multi-label Data Streams». En: (), pág. 16.

[10] Jesse Read y col. «Classifier chains for multi-label classification». En: *Mach. Learn.* 85.3 (2011), págs. 333-359.

[11] Jesse Read y col. «Scalable and efficient multi-label classification for evolving data streams». En: *Machine Learning* 88.1 (1 de jul. de 2012), págs. 243-272. ISSN: 1573-0565. DOI: 10.1007/s10994-012-5279-6. URL: https://doi.org/10.1007/s10994-012-5279-6 (visitado 17-06-2020).

[12] Ricardo Sousa y João Gama. «Multi-label classification from high-speed data streams with adaptive model rules and random rules». En: *Progress in Artificial Intelligence* (2018).

[13] Grigorios Tsoumakas y Ioannis Katakis. «Multi-Label Classification». En: *Int. J. Data Warehouse. Min.* 3.3 (2007), págs. 1-13.

[14] Min-Ling Zhang y Kun Zhang. «Multi-label learning by exploiting label dependency». En: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10*. 2010.

[15] Min-Ling Zhang y Zhi-Hua Zhou. «A Review On Multi-Label Learning Algorithms». En: *IEEE Trans. Knowl. Data Eng.* 26 (2014), págs. 1819-1837.