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.

AGRADECIMIENTOS

Acá, agradezco...

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 [5]. 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 [9].

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* [7]. 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. Clasificación de flujos de datos multi etiquetados

Datos multietiquetados.

Aprendizaje incremental ?

Streamings. Caracteristicas esenciales (potencialmente infinita, limite de espacio en memoria y de tiempo, etc).

Caracteristicas de una colección multietiquetada (densidad, cardinalidad,etc).

1.3. Motivación

Ante la necesidad de hacer frente a un contexto global de generación masiva de datos y a un ritmo cada vez mayor, 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 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 limitación en el espacio de almacenamiento. Si bien existen métodos de clasificación para flujos continuos que han dado resultados satisfactorios, aún es un campo poco abordado y se hace necesario reproducir los experimentos realizados y fortalecer las técnicas y herramientas actuales para llevar adelante estudios precisos y pormenorizados.

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 ca-

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

racterísticas de un flujo continuo de datos. Uno de los enfoques abordados es convertir las colecciones existentes en flujos que arriban a lo largo del tiempo y en cantidades predefinidas. De esta manera los algoritmos pueden 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 area. Estos son capaces de generar datos cercanos a los de colecciones del mundo real [6] y brindan la posibilidad de realizar estudios certeros de algoritmos de clasificación [8]. Pero 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 podrá 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.

explicar enfoque de clasificación por ensambles.

1.4. Objetivos

Acá van los objetivos.

1.5. Aportes

Acá van los aportes.

1.6. Organización del Trabajo

Acá va la descripción de las distintas secciones del trabajo.

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] 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.

[6] Jesse Read, Bernhard Pfahringer y Geo Holmes. «Generating Synthetic Multi-label Data Streams». En: *2009* (), pág. 16.

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

[8] 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).

[9] 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).