Modelos para Aprendizaje Automático en Tiempo Real sobre Entornos de Big Data

Banchero Santiago[1], Fernandez Juan M.[1], Tonin Monzón Francisco[1], Giordano Luis A.[1]
Marrone Agustín H.[1], Lulic Maximiliano[1], Tolosa Gabriel H.[1,2]
{sbanchero, jmfernandez, ftonin, agiordano, amarrone, mlulic, tolosoft}@unlu.edu.ar
[1]Departamento de Ciencias Básicas, Universidad Nacional de Luján
[2]CIDETIC, Universidad Nacional de Luján

Resumen

En la actualidad existen incontables fuentes de información en tiempo real que provienen de redes de sensores, plataformas de observación del tiempo, mediciones de gases, observación de la tierra desde plataformas satelitales, ciudades inteligentes, entre un sin número de instrumentos que censan y transmiten datos. A su vez, hay una creciente demanda por el desarrollo de herramientas que permitan extraer conocimiento a partir de esos grandes repositorios de datos. El aprendizaje automático es un área de la inteligencia artificial, donde sus métodos contribuyen en el proceso de descubrimiento de conocimiento para la toma de decisiones inteligentes. Las demandas para la extracción de conocimiento en entornos de Big Data han acrecentado el interés por la utilización de técnicas tradicionales de aprendizaje automático en distintos problemas de repositorios masivos y entornos de flujos (o *streaming*) de datos donde muchas veces no es posible su almacenamiento, pero se requiere tomar decisiones en tiempo real.

Contexto

Este proyecto es el comienzo de una nueva línea de investigación del Departamento de Ciencias Básicas (UNLu) que tiene como principal aspiración profundizar los conocimientos sobre métodos actuales de aprendizaje de máquina como herramienta para el descubrimiento de conocimiento en problemas de Big Data sobre *streaming* de datos de diversas naturaleza.

Introducción

En la actualidad existen muchas aplicaciones que hacen un uso intensivo de datos, haciendo que el volumen y la complejidad de estos crezcan rápidamente. Los motores de búsqueda, redes sociales, e-ciencia (por ejemplo: genómica, meteorología y salud) y financieras (por ejemplo: banca y megatiendas entre otros) son algunas de sus aplicaciones [1, 15]. Esta problemática se conoce como el problema de Big Data [19].

Big Data se caracteriza principalmente por tres aspectos: (a) los datos son numerosos, (b) los datos no pueden ser categorizados en bases de datos relacionales regulares, y (c) los datos son generados, capturados y procesados muy rápidamente [13]. Si bien el volumen es y será un desafío significativo del Big Data, se debe prestar mucha atención a todas las dimensiones del problema: Volumen, Variedad y Velocidad (conocidos como las 3Vs) [6].

El concepto de *streaming* en Big Data tiene algunas características distintivas, en estos sistemas los datos se reciben como una secuencia continua, infinita, rápida, en ráfagas, impredecible y que varía en el tiempo [8]. El monitoreo (por ejemplo: tráfi-

co de red, redes de sensores, cuidado de la salud, etc.), seguimientos de *clics* en la web, transacciones financieras, detección de fraudes e intrusiones son algunas aplicaciones de *streaming* de Big Data [5]. Todos estos productores de datos que generan el *streaming* a menudo se encuentran distribuidos y con capacidades de procesamiento y memoria limitados.

En tareas de extracción de conocimiento dos pasos muy importantes son la selección de *features* (o atributos) y las tareas de mining de datos [17]. Los entornos de *streaming* de datos proponen nuevos retos con respecto a estas etapas. Una característica importante en selección de *features* es la habilidad para manejar grandes volúmenes de datos [23, 26]. Gran parte de las publicaciones existentes en la Web arriban como *streaming* (documentos, imágenes, contenidos multimedia, etc.), detectar un subconjunto de *features* útiles en estos flujos de datos en una tarea compleja debido a limitaciones de memoria, tiempos de respuesta, etc. [12, 29, 28].

Además del problema de selección de *features*, hay cuestiones enmarcadas dentro de las tareas de *stream mining* para extracción de conocimiento. La problemática aquí radica en que los patrones de datos evolucionan continuamente y se torna necesario diseñar algoritmos de minería para tener en cuenta los cambios en la estructura subyacente del *streaming* de datos [3, 2, 27]. Incluso la distribución subyacente puede cambiar en el tiempo, lo que genera que algunos modelos ya no sigan siendo válidos. Estos aspectos hacen que las soluciones de los problemas sean aún más difíciles desde un punto de vista algorítmico y computacional [7].

En este trabajo se proponen diversas líneas de investigación sobre los temas mencionados, con aplicaciones en flujos de datos y problemas reales. Se abordan tanto problemas de mejoras de rendimiento ante distintos niveles de exigencia de precisión como también la escalabilidad de las diferentes aplicaciones a datos reales.

Líneas de I+D

En este proyecto se llevan adelante lineas de I+D relacionadas principalmente con el análisis, adaptación y prueba de algoritmos de aprendizaje automático en entornos de *Streaming* de datos. Las principales líneas de trabajo en la actualidad proyecto son:

- Gestión de cache de consultas. Se busca determinar qué algoritmos de aprendizaje - en entornos de streaming - permiten lograr una mejor performance de clasificación en vías de optimizar la gestión de un cache de consultas.
- Algoritmos de clasificación multilabel. El objetivo de esta línea de trabajo es explorar los diferentes abordajes en un proceso de aprendizaje multi-etiquetas en ambientes de *streaming* de datos.
- Data Stream Clustering. Se busca desarrollar una metodología de detección de novedades o valores atípicos a través de algoritmos de agrupamiento. Analizando flujos de datos sintéticos y provenientes de redes sociales.

A continuación se hace una descripción somera de estas líneas de I+D.

a. Gestión de cache de consultas

Los árboles de decisión corresponden al aprendizaje supervisado y son ampliamente utilizados en problemas de clasificación. Estos algoritmos intentan, a partir de las instancias vistas, generar hipótesis con las cuales hacer predicciones de futuras instancias [18].

En aprendizaje sobre flujo de datos (*stream learning*), en general no es necesario computar estadísticas sobre todo el pasado, siendo suficiente con hacerlo sobre el pasado reciente [9]. Una de las formas mas clásicas y simples de mantener los ejemplos correspondientes a ese pasado es almacenar solo una ventana de instancias.

Los árboles de decisión adaptativos, o *Hoeffding Adaptive Tree* (HAT) [3], son una variante de *Hoeffding Tree* que utilizan ventanas deslizantes para

mantener ajustado el árbol, sin embargo no requiere que el usuario le especifique el tamaño de ventana a utilizar. Esto se debe a que el tamaño de ventana óptimo se calcula individualmente para cada nodo, utilizando detectores de cambios y estimadores llamados ADWIN [4].

Los resultados preliminares de aplicar HAT en el dominio de gestión de caché de consultas en motores de búsqueda han sido muy alentadores [25]. Como objetivo principal, se propuso evaluar la performance de un árbol de decisión adaptativo (HAT) para aplicar al diseño de una política de admisión para un motor de búsqueda web que recibe (y procesa) consultas en modo *streaming*. Se ha trabajado modelando la admisión como un problema de clasificación binario, intentando capturar los cambios de concepto en el tiempo, manteniendo un modelo siempre ajustado. Una vez ajustado el modelo, se integró el mecanismo de decisión como política de admisión y se evaluó la performance del caché de resultados, comparando el modelo resultante de utilizar árboles de decisión adaptativos con algoritmos de clasificación tradicionales. Siendo, de acuerdo a nuestro conocimiento, la primera vez que se propone un árbol de decisión adaptativo para la detección de términos de búsqueda frecuentes en motores de búsqueda, en los experimentos realizados, se ha observado un incremento del rendimiento del 18 % en comparación con la utilización de técnicas de clasificación tradicionales.

Los próximos pasos para esta línea de investigación están orientados a evaluar alternativas de clasificación a HAT. En este sentido se planea abordar dos alternativas, una de menor complejidad que tiene que ver con Naïve Bayes [8] y una de mayor complejidad asociada a un algoritmo de ensamble como *Random Forest* [11]. El objetivo de este trabajo es identificar posibles cotas de exactitud que puedan existir a la hora de escoger un algoritmo de aprendizaje automático sobre *stream* data que decida si una consulta es única o se repetirá en el corto plazo.

b. Algoritmos de clasificación multi-etiquetas

La clasificación multi-etiquetas es un nuevo 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 [10].

La mayor parte de los trabajos de investigación en este campo han sido realizados en contextos de aprendizaje por *batch* [8]; sin embargo, los ambientes de flujo continuo de datos (o *streaming*) presentan nuevos desafíos 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 [22].

Una propuesta para esta linea de investigación sugiere aplicar algoritmos de clasificación multi-etiquetas a colecciones estructuradas y no estructuradas, combinando estos algoritmos con técnicas de procesamiento de lenguaje natural sobre la colección no estructurada. A su vez, por último, se proponen abordar 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.

c. Data Stream Clustering

Data Stream Clustering puede ser planteado como un problema tradicional de agrupamiento donde existen un conjunto finito (pero desconocido) de categorías - o *clusters* - que permiten describir la estructura de un conjunto de datos. El fundamento detrás de los algoritmos de *clustering* es que los objetos dentro de un *cluster* son más parecidos entre sí que los objetos que pertenecen a un *cluster* diferente [21, 16]. En este contexto, varios algoritmos de *clustering* sobre data *stream* han sido propuestos para realizar aprendizaje no supervisado.

El agrupamiento sobre un *stream* de datos requiere que un proceso sea capaz de agrupar objetos continuamente dentro de las restricciones de memoria y tiempo [8] que son las dos limitantes del problema. Los algoritmos de agrupamiento en flu-

jos de datos deben ser capaces de cumplir con los siguientes requerimientos [24, 3]:

1. Proveer oportunamente resultados mediante el procesamiento rápido e incremental de objetos;

2. adaptarse con rapidez a los cambios en la dinámica de los datos, es decir, un algoritmo debe detectar nuevos clusters que van apareciendo y otros que desaparecen;

3. escalar a la cantidad de objetos que constantemente arriban;

4. proveer un modelo de representación que no solo es compacto sino que además no crece con el número de objetos procesados;

5. capacidad para detectar *outliers*;

6. tratar con diferentes tipos de datos.

En el mundo real, pueden aparecer muchos datos atípicos o que son ruido debido a situaciones atribuidas a fallas de sensores, problemas con las conexiones, etc. Los algoritmos de agrupamiento deben ser resistentes a los valores atípicos. En un entorno de *streaming* de datos, esto es un gran desafío ya que tanto los *clusters* como los valores atípicos evolucionan con el tiempo. En otras palabras, es difícil saber ante el arribo de un nuevo dato si este se trata de un valor atípico o es el primer miembro de un nuevo grupo [30, 20, 14].

En esta línea de investigación se propone aplicar algoritmos de agrupamiento sobre flujos de datos para la detección de valores atípicos en diferentes niveles de dificultad. Inicialmente trabajará en la simulación de varios escenarios de flujos construidos de manera sintética con el objetivo de evaluar rendimiento de los algoritmos ante situaciones con variaciones en la distribución de los datos, cambios de conceptos e incorporación de ruido. Otra linea de trabajo será evaluar la detección de valores atípicos en flujos del mundo real, como por ejemplo Twitter.

Resultados y Objetivos

El objetivo principal de la propuesta es estudiar, desarrollar, aplicar, validar y transferir modelos, algoritmos y técnicas que permitan construir herramientas y/o arquitecturas para abordar algunas de las problemáticas relacionadas con el tratamiento de información masiva utilizando algoritmos de aprendizaje automático de Big Data para dar respuestas en tiempo real. Se propone profundizar sobre el estado del arte y definir, analizar y evaluar nuevos enfoques sobre aprendizaje automático a partir de *streaming* de datos. En particular se estudiarán las siguientes líneas principales:

1. Estrategias de gestión *streaming* de datos masivos para determinar las mejores herramientas para extracción de features y resolución de los problemas clásicos de ETL en el contexto del real-time.

2. Evaluar la escalabilidad de los algoritmos tradicionales del área de aprendizaje automático a problemas de respuestas en tiempo real sobre *streaming* de datos masivos en diferentes dominios.

3. Elaborar metodologías para el desarrollo de modelos en linea para toma de decisiones a partir de fuentes de información heterogénea.

Formación de Recursos Humanos

Este proyecto brinda un marco para que algunos docentes auxiliares y estudiantes lleven a cabo tareas de investigación y se desarrollen en el ámbito académico. Hasta la fecha se ha realizado un trabajo final correspondiente a la Lic. en Sistemas de Información (UNLu), se está dirigiendo uno más. Se espera dirigir al menos dos por año hasta la finalización del proyecto.

Referencias

[1] Aggarwal, C. C., Ashish, N., and Sheth, A. The internet of things: A survey from the data-centric perspective. In *Managing and mining sensor data*. Springer, 2013, pp. 383–428.

[2] Baldominos, A., Albacete, E., Saez, Y., and Isasi, P. A scalable machine learning online service for big data real-time analysis. In *Computational Intelligence in Big Data (CIBD), 2014 IEEE Symposium on* (2014), IEEE, pp. 1–8. 00017.

[3] Bifet, A. Adaptive stream mining: Pattern learning and mining from evolving data streams. In *Proceedings of the 2010 conference on adaptive stream mining: Pattern learning and mining from evolving data streams* (2010), Ios Press, pp. 1 212.

[4] Bifet, A., and Gavalda, R. Learning from time-changing data with adaptive windowing. In *Proceedings of the 2007 SIAM international conference on data mining* (2007), SIAM, pp. 443–448.

[5] Bifet, A., and Morales, G. D. F. Big data stream learning with samoa. In *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on* (2014), IEEE, pp. 1199–1202.

[6] Chen, M., Mao, S., and Liu, Y. Big data: a survey. *Mobile Networks and Applications 19*, 2 (2014), 171–209. 00324.

[7] Gaber, M. M., Zaslavsky, A., and Krishnaswamy, S. Mining data streams: a review. *ACM Sigmod Record 34*, 2 (2005), 18–26.

[8] Gama, J. *Knowledge discovery from data streams*. CRC Press, 2010.

[9] Gama, J., and Gaber, M. M. *Learning from data streams: processing techniques in sensor networks*. Springer, 2007.

[10] Gibaja, E., and Ventura, S. A tutorial on multilabel learning. *ACM Computing Surveys (CSUR) 47*, 3 (2015), 52.

[11] Gomes, H. M., Bifet, A., Read, J., Bardal, J. P., Enembreck, F., Pfharinger, B., Holmes, G., and Abdessalem, T. Adaptive random forests for evolving data stream classification. *Machine Learning 106*, 9-10 (2017), 1469–1495.

[12] Huang, H., Yoo, S., and Kasiviswanathan, S. P. Unsupervised feature selection on data streams. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (2015), pp. 1031–1040.

[13] Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Mahmoud Ali, W. K., Alam, M., Shiraz, M., and Gani, A. Big data: survey, technologies, opportunities, and challenges. *The Scientific World Journal 2014* (2014). 00028.

[14] Mansalis, S., Ntoutsi, E., Pelekis, N., and Theodoridis, Y. An evaluation of data stream clustering algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal 11*, 4 (2018), 167–187.

[15] Marz, N., and Warren, J. *Big Data: Principles and best practices of scalable realtime data systems*. Manning Publications Co., 2015.

[16] Nguyen, H.-L., Woon, Y.-K., and Ng, W.-K. A survey on data stream clustering and classification. *Knowledge and information systems 45*, 3 (2015), 535–569.

[17] Pruengkarn, R., Wong, K., and Fung, C. A review of data mining techniques and applications. *Journal of Advanced Computational Intelligence and Intelligent Informatics 21*, 1 (2017), 31–48.

[18] Quinlan, J. R. Induction of decision trees. *Machine learning 1*, 1 (1986), 81–106.

[19] Safaei, A. A. Real-time processing of streaming big data. *Real-Time Systems 53*, 1 (2017), 1–44. 00004.

[20] Shao, X., Zhang, M., and Meng, J. Data stream clustering and outlier detection algorithm based on shared nearest neighbor density. In *2018 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)* (2018), IEEE, pp. 279–282.

[21] Silva, J. A., Faria, E. R., Barros, R. C., Hruschka, E. R., Carvalho, A. C. d., and Gama, J. Data stream clustering: A survey. *ACM Computing Surveys (CSUR) 46*, 1 (2013), 1–31.

[22] Sousa, R., and Gama, J. Multi-label classification from high-speed data streams with adaptive model rules and random rules. *Progress in Artificial Intelligence 7*, 3 (2018), 177–187.

[23] Tang, J., Alelyani, S., and Liu, H. Feature selection for classification: A review. *Data Classification: Algorithms and Applications* (2014), 37.

[24] Tasoulis, D. K., Adams, N. M., and Hand, D. J. Unsupervised clustering in streaming data. In *Sixth IEEE International Conference on Data Mining-Workshops (ICDMW'06)* (2006), IEEE, pp. 638–642.

[25] Tonin Monzón, F., Banchero, S., and Tolosa, G. H. Árboles de decisión adaptativos en políticas de admisión a caché. In *IV Simposio Argentino de GRANdes DAtos (AGRANDA 2018)-JAIIO 47 (CABA, 2018)* (2018).

[26] Wang, J., Zhao, P., Hoi, S. C., and Jin, R. Online feature selection and its applications. *IEEE Transactions on Knowledge and Data Engineering 26*, 3 (2014), 698–710.

[27] Wang, L. Machine learning in big data. *International Journal of Advances in Applied Sciences 4*, 4 (2016), 117–123. 00048.

[28] Wu, X., Yu, K., Wang, H., and Ding, W. Online streaming feature selection. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (2010), Citeseer, pp. 1159–1166.

[29] Yin, C., Feng, L., Ma, L., Wang, J., Yin, Z., and Kim, J.-U. A feature selection algorithm of dynamic data-stream based on hoeffding inequality. In *2015 4th International Conference on Advanced Information Technology and Sensor Application (AITS)* (2015), IEEE, pp. 92–95.

[30] Yin, C., Zhang, S., Yin, Z., and Wang, J. Anomaly detection model based on data stream clustering. *Cluster Computing* (2017), 1–10.