Algoritmos de aprendizaje automático para respuestas en tiempo real sobre entornos masivos de datos

Banchero Santiago, Tolosa Gabriel H., Tonin Monzón Francisco,
Fernandez Juan M., Paz Soldan Carlos, Giordano Luis A.,
Marrone Agustín H.
{sbanchero, tolosoft, ftonin, jmfernandez, cpazsoldan, agiordano, amarrone}@unlu.edu.ar
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 para poder 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 ha acrecentado el interés por la utilización de técnicas tradicionales de aprendizaje automático en distintos problemas de repositorios masivos de datos y también en entornos de flujos (o *streaming*) de datos donde muchas veces no es posible su almacenamiento, pero se requiere tomar decisiones en tiempo real conforme se leen los datos.

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), financieros (por ejemplo: banca y megatiendas entre otros) son algunos ejemplos de tales aplicaciones [1, 11]. Esta problemática se conoce como el problema de Big Data [13].

Big Data se caracteriza 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 [10]. Si bien el volumen es y será, un desafío significativo del Big Data se debe prestar mucha atención en todas las dimensiones del problema: Volumen, Variedad y Velocidad (conocidos como las 3Vs) [6].

El concepto de *streaming* en Big Data tiene algunas características notables, 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: tráfico 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 [12]. Estos problemas en entornos de *streaming* de datos siguen siendo un gran desafío. Una característica importante en selección de *features* es la habilidad para manejar grandes volúmenes de datos [14, 15]. 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. [9, 18, 17].

Además del problema de selección de *features*, hay cuestiones enmarcadas dentro de las tareas de *streaming* 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, 16]. Incluso la distribución subyacente puede cambiar en el tiempo lo que genera que algunos modelos ya no sigan siendo válidos. Esto hace 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 los diferentes abordajes a datos reales.

Líneas de I+D

En este proyecto se inician lineas de I+D relacionadas principalmente con el análisis y adaptación de algoritmos de aprendizaje automático en entornos de *Streaming* de datos. Puntualmente se está trabajando con árboles de decisión adaptativos en aplicaciones de cache de consultas sobre motores de búsqueda. Por otro lado, se está trabajando en profundizar los conocimientos en selección de atributos sobre *streaming* de datos. Por último, se están analizando diferentes opciones de topologías con Storm Apache[1] para la resolución de problemas de ETL. A continuación se hace una descripción somera de las líneas de I+D.

a. Árboles de decisión adaptativos

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.

Nuestro siguiente paso es realizar experimentos en otros dominios como, por ejemplo: análisis de sentimientos en redes sociales. De esta manera esperamos cuantificar la respuesta de éste método de extración de conocimiento en contextos sometidos a constantes cambios de conceptos.

b. Selección de atributos

Enormes fuentes de datos se generan continuamente a partir de fuentes tales como redes sociales, difusión de noticias, etc., y típicamente estos datos se encuentran en espacios de alta dimensión (como el espacio de vocabulario de un idioma). La selección de atributos, o *Feature Selection* (FS), para

[1] `http://storm.apache.org/`

descubrimiento de conocimiento, ha representado un gran desafío durante los últimos años tanto en estadística como en problemas de aprendizaje automático [17]. En el contexto de *streaming* de datos de gran volumen, detectar un subconjunto de *features* que sean relevantes es un problema muy difícil de resolver por las siguientes razones:

1. El *stream* de datos puede ser infinito, por lo que cualquier algoritmo que trabaje *off-line* que intente almacenar toda la secuencia para el análisis se quedará sin memoria.
2. La importancia de los atributos cambia dinámicamente con el tiempo debido a la volatilidad de un concepto, un atributo importante pueden volverse insignificantes y viceversa.
3. Para varias aplicaciones en línea, es importante obtener el subconjunto de características en tiempo casi real.

Esta línea de investigación es complementaria de la anterior y se propone analizar la precisión en métodos supervisados para selección de *features* en fuentes de datos no estructuradas provenientes de redes sociales. Este abordaje va a permitir trabajar la dinámica de los cambios de conceptos en *streaming* de datos en tareas de clasificación.

c. Topologías Storm

En esta línea de investigación se propone trabajar con herramientas de *Stream Processing Engine* (SPE) para probar los diferentes algoritmos de aprendizaje automático (como HAT) y las diferentes estrategias de selección de atributos. Un SPE es un framework que tiene por objetivo abordar el desafío de procesar grandes volúmenes de datos, en tiempo real y sin requerir el uso de código específico. Sobre los SPE es posible implementar algoritmos de *machine learning* para extraer conocimiento de los *streaming* de datos.

La idea de utilizar herramientas como Apache Storm es poder definir topologías de procesamiento de manera ágil para gestión de estadísticas necesarias para las etapas de selección de *features* y en tareas de aprendizaje supervisado y no supervisado.

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. Elaboración de 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. Actualmente, se están dirigiendo dos trabajos finales correspondientes a la Lic. en Sistemas de Información (UNLu). Se espera dirigir al menos dos estudiantes más por año y presentar dos candidatos a becas de investigación.

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] Huang, H., Yoo, S., and Prasad, S. Unsupervised Feature Selection on Data Streams. 1031–1040.

[10] 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.

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

[12] 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.

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

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

[15] 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.

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

[17] 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.

[18] Zhichao, Y., and Chunyong Yin, L. F. A Feature Selection Algorithm of Dynamic Data-Stream Based on Hoeffding Inequality. 92–95.