Impacto del Machine Learning en la Seguridad, Ciencia y Big Data

Crisly González Sánchez
Escuela Ingeniería en Computación
Instituto Tecnológico de Costa Rica
Alajuela,San Carlos
gonzalezcrisly@gmail.com

***Resumen*—El aprendizaje de las máquinas(ML) beneficia las actividades que realiza actualmente el ser humano desde el campo de la medicina, entretenimiento, seguridad y monitoreo de la red, la investigación hace una revisión exhaustiva de en los diferentes campos anteriormente mencionados para mostrar usos de la tecnología machine learning, además se abarca la relación entre machine learning, las redes neuronales y big data.**

Keywords*—*Machine Learning(ML), Big Data, Montoreño Red, Modelos de datos,Algoritmos de predicción y aprendizaje.

I. Introducción

Machine Learning(ML) es una disciplina del ámbito de la Inteligencia Artificial, un ejemplo para demostrar su funcionamiento es la creación de sistemas que aprenden automáticamente, cabe destacar que ML se fortalece de la identificación de patrones en grandes cantidades de información, de esta forma los sistemas aprendan de acuerdo a probabilidades de cual es la siguiente acción a realizar. CleaverData (s.f.)

El aprendizaje de las máquinas está estrechamente relacionado con la robótica, lo que nos demuestra que no es un concepto nuevo, desde hace muchos años se ha estado trabajando los distintos sectores de la industria de las tecnologías en el mejoramiento de máquinas que faciliten el trabajo a los usuarios y por consiguiente robots que aprendan a tomar decisiones como los humanos, esto nos permite entrar en el campo de la minería de datos (Big Data) ¿Cómo lograr el aprendizaje de las máquinas por medio de análisis de datos?¿Es precisa la información que aprende una máquina? ¿Algoritmos que favorecen el aprendizaje de las máquinas? todas estas preguntas serán abordadas en el desarrollo de la investigación.

Abordaremos algoritmos de seguridad utilizados para el monitoreo de las redes mediante ML, la precisión en los resultados para mostrar alguna anomalía es compleja ya que la información procesada se realiza mediante modelos diseñados para analizar un comportamiento y los picos en red pueden ser variados, dejando un hueco en el área de seguridad para dar resultados precisos, sin embargo lo investigado nos da un preámbulo de hasta dónde podemos llegar con el aprendizaje automático. (Bhamare y cols., 2016)

Actualmente muchas empresas como Amazon, Google e incluso Netflix invierten en departamentos de científicos de datos o colaboradores a nivel mundial que quieran analizar información contenida por sus clientes y así mostrar una mejor experiencia, en la sección II de trabajos relacionados se hablará del caso de éxito de Netflix en el tema Machine Learning. Por otra parte desde el ámbito tecnológico y herramientas libres que se pueden utilizar, la empresa de tecnología más grande a nivel mundial pone a disposición servicios en la nube como Google Cloud Vision, Speech, Natural Language, entre otros GoogleCommunity (s.f.).

Finalmente es importante destacar que Machine Learning hace uso de redes neuronales, árboles de decisión y regresiones, esto podemos ejemplificar con los demo de los juegos que generalmente están en internet explicando cómo jugar, en el que las soluciones son un conjunto de nodos que deben ser visitados y guardar las llamadas que son exitosa, haciendo que la máquina grabe las jugadas hasta hacer el demo completo mediante los casos de éxito Vallarino (s.f.).

II. Trabajos Relacionados

Machine Learning ha sido aprovechado para automatizar procesos que anteriormente requerían trabajo dedicado de personas en disposición de estar analizando comportamientos, para esto la tecnología permite crear patrones de comportamiento para ser analizado por medio de algoritmos, lo que más precisa la evaluación de comportamientos y además permite procesar muchos datos sin trabajo humano, únicamente con el aprendizaje que tiene la máquina.

(Amatriain, 2013)Empresas como Netflix han experimentado la necesidad de implementar en su plataforma algoritmos de predicción que brindarán al cliente recomendaciones según sus búsquedas, la empresa recalca la importancia de utilizar los datos a bien de sí mismos para brindar una excelente navegación a los clientes, además indican que el éxito de un negocio se ve por la forma en la cual pueden tratar los datos de su cliente para brindarles mejor servicio, es por esto que incluso otras compañías tienen sus científicos de datos pendientes para optimizar y mejorar procesos.

Mediante la tecnología de aprendizaje de las máquinas podemos controlar la seguridad de distintas aplicaciones que manejan gran cantidad de información sensible, en (Lopez y Cadavid, 2016) se propone la implementación de algoritmos para detectar el malware que usualmente se insertan en los sistemas para robo de información, los escritores proponen los siguientes algoritmos usualmente utilizados en sistemas de seguridad (Naive Bayes, ensacado, KNeighbors, Super Vector Machines (SVM), Descenso de gradiente estocástico (SGD)

y Árbol de decisión) determinando que el mejor según las pruebas realizadas es KNeighbors.

(Bhamare y cols., 2016)Los usuarios de los servicios en la nube están bajo constante temor de pérdida de datos, amenazas de seguridad y problemas de disponibilidad. Recientemente, los métodos basados en el aprendizaje para aplicaciones de seguridad están ganando popularidad en la literatura con las ventajas de las técnicas de aprendizaje automático. Sin embargo, el principal reto en estos métodos es la obtención de datos en tiempo real e imparcial. Para la evaluación del rendimiento al implementar algoritmos de detección de patrones en la red se utilizó los anteriormente mencionados, sin embargo, se determina que aunque los comportamientos pueden ser distintos y esto deja un vacío en su efectividad, no obstante disminuyen los ataques que puedan realizarse al determinar comportamientos a tiempo, sin embargo aún sigue faltando una perfección en esta tecnología para ser aplicada en la seguridad de la nube.

(Peixoto y cols., 2016)Durante la investigación se determina su relación con Big Data, ambas tecnologías son utilizadas en conjuntos para procesar obtener datos, procesarlos y dar resultados mediante modelos de regresión lineal que es la técnica más conocida para modelar datos. Es importante mencionar que hay bases de datos distintas para realizar el almacenamiento de las máquinas con ML con Big Data, los lectores mencionan la base de conocimiento de la ontología Abox(orientada a las instancias de los objetos) + Tbox(orientadas a objetos) utilizado para representar el conocimiento en la Sistema de clasificación se aprende automáticamente de grandes volúmenes de datos a través de técnicas altamente escalables de Aprendizaje Grandes Tecnologías de Datos.

Finalmente analizamos con los lectores (Baeza-Yates y cols., 2015) la predicción de las aplicaciones móviles que serán instaladas por el usuario, cabe destacar que las aplicaciones móviles son la más inestables con respecto a su permanencia en el móvil, ya que al haber diversidad de aplicaciones los usuarios las reemplazan en todo momento. Basado en la premisa anterior los investigadores idearon un mecanismo para determinar cuándo se va desinstalar la aplicación de acuerdo a las variables tiempo de uso en un determinado lapso establecido por ellos, cantidad de descargas y popularidad en la tienda Play Store. Se realizaron investigaciones con una cantidad de datos y se logró una precisión de un 10,5 %.

III. Conclusión

En (Somvanshi y Chavan, 2016)se utilizó el algoritmo ID3 que es muy fácil y capaz de clasificar grandes conjuntos de datos. Mostrándonos la relación estrecha que existe entre la minería de datos y el aprendizaje de las máquinas. El estudio muestra cómo el árbol de decisiones(Técnica de Machine Learning) se puede tomar y ser utilizado de manera eficiente con datos discretos.

En (Kim y cols., 2017) se analizan los siguientes algoritmos Métodos de aprendizaje de la máquina como árbol de decisión, Bayes naïve, apriori, KNN, K-means, SVM.Enfoque de redes neuronales como CNN, RNN, DBM. Con el objetivo de analizar el filtrado semántico.

En (Lopez y Cadavid, 2016; Bhamare y cols., 2016) se realizan análisis en algoritmos para el monitoreo de sistemas en la nube que son vulnerables a malware y deben crearse mecanismos para analizar patrones de comportamiento en los paquetes de red, lo cual determina la deficiencia de control que se puede ejercer mediante Machine Learning ya que los comportamientos son variantes y pueden presentarse una única vez, por lo tanto el control y aprendizaje de la máquina se puede enfrentar a una situación no esperada vulnerando la información.

En (Fageeri y cols., 2017) los escritores realizaron una prueba de campo de ML para beneficio de la salud médica,se tomaron un conjunto de datos de 1530 instancias y 10 atributos. Se seleccionaron cuatro categorías como etiquetas de clasificación: "Miopía", "Hipermetropía", .ªstigmatismo de hipermetropíaz .ªstigmatismo de miopía". Posteriormente se usaron tres técnicas de aprendizaje automático para predecir la gravedad del ojo, se utilizaron los siguientes algoritmos Bayesian Naïve, SVM y el clasificador de árboles de decisión J48 y se determinó que el mejor era J48.

En (Jadhav y cols., 2016) se realizó un método integrado de clasificación de información de texto semántico, la imagen y en la información de concordancia de objetos. Utilizando las herramientas OpenCV, WordNet se identificó el problema asociado con los sistemas de búsqueda tradicionales junto con la información semántica que determinando que las imágenes no se consideran para extraer la información semántica.

En (Waegel y Kontostathis, 2006)se analiza la herramienta TextMOLE: Text Mining Operations Library and Environment que es un motor de búsqueda y indexación avanzada: analiza un conjunto de datos, extrae términos relevantes y permite al usuario ejecutar consultas en contra de los datos. La herramienta está diseñada para analizar rápidamente un corpus de documentos y determinar qué parámetros proporcionarán un rendimiento máximo de recuperación mediante modelos de regresión.

Referencias

Amatriain, X. (2013). Beyond data: from user information to business value through personalized recommendations and consumer science. En *Proceedings of the 22nd acm international conference on conference on information & knowledge management* (pp. 2201–2208). New York, NY, USA: ACM. Descargado de http://doi.acm.org/10.1145/2505515.2514701 doi: 10.1145/2505515.2514701

Baeza-Yates, R., Jiang, D., Silvestri, F., y Harrison, B. (2015). Predicting the next app that you are going to use. En *Proceedings of the eighth acm international conference on web search and data mining* (pp. 285–294). New York, NY, USA: ACM. Descargado de http://doi.acm.org/10.1145/2684822.2685302 doi: 10.1145/2684822.2685302

Bhamare, D., Salman, T., Samaka, M., Erbad, A., y Jain, R. (2016, Dec). Feasibility of supervised machine learning for cloud security. En *2016 international conference on information science and security (iciss)* (p. 1-5). doi: 10.1109/ICISSEC.2016.7885853

CleaverData. (s.f.). *¿qué es machine learning?* Descargado de http://cleverdata.io/que-es-machine-learning-bi

Fageeri, S. O., Ahmed, S. M. M., Almubarak, S. A., y Mu'azu, A. A. (2017, Jan). Eye refractive error classification using machine learning techniques. En *2017 international conference on communication, control, computing and electronics engineering (icccceee)* (p. 1-6). doi: 10.1109/ICCCCEE.2017.7867660

GoogleCommunity. (s.f.). *Cloud machine learning services.* Descargado de https://cloud.google.com/products/machine-learning/

Jadhav, P. A., Chatur, P. N., y Wagh, K. P. (2016, Feb). Integrating performance of web search engine with machine learning approach. En *2016 2nd international conference on advances in electrical, electronics, information, communication and bio-informatics (aeeicb)* (p. 519-524). doi: 10.1109/AEEICB.2016.7538344

Kim, M., Kang, H., Kwon, S., Lee, Y., Kim, K., y Pyo, C. S. (2017, Feb). Augmented ontology by handshaking with machine learning. En *2017 19th international conference on advanced communication technology (icact)* (p. 740-743). doi: 10.23919/ICACT.2017.7890191

Lopez, C. C. U., y Cadavid, A. N. (2016, April). Machine learning classifiers for android malware analysis. En *2016 ieee colombian conference on communications and computing (colcom)* (p. 1-6). doi: 10.1109/ColComCon.2016.7516385

Peixoto, R., Hassan, T., Cruz, C., Bertaux, A., y Silva, N. (2016). An unsupervised classification process for large datasets using web reasoning. En *Proceedings of the international workshop on semantic big data* (pp. 9:1–9:6). New York, NY, USA: ACM. Descargado de http://doi.acm.org/10.1145/2928294.2928301 doi: 10.1145/2928294.2928301

Somvanshi, M., y Chavan, P. (2016, Aug). A review of machine learning techniques using decision tree and support vector machine. En *2016 international conference on computing communication control and automation (iccubea)* (p. 1-7). doi: 10.1109/ICCUBEA.2016.7860040

Vallarino, D. (s.f.). *La teoría de juegos en la construcción de machine learning.* Descargado de https://www.forbes.com.mx/la-teoria-de-juegos-en-la-construccion-de-machine-learning/

Waegel, D. B., y Kontostathis, A. (2006, marzo). Textmole: Text mining operations library and environment. *SIGCSE Bull.*, *38*(1), 553–557. Descargado de http://doi.acm.org/10.1145/1124706.1121511 doi: 10.1145/1124706.1121511