IV Conferencia Científica Internacional
Universidad de las Ciencias Informáticas

Temática: Ciberseguridad

Algoritmos de detección de anomalías con redes profundas. Revisión para detección de fraudes bancarios

Anomaly detection algorithms with deep networks. Review for Bank Fraud Detection

David Ameijeiras Sánchez[1*], Héctor R. González Diez[2]

[1]Universidad de las Ciencias Informáticas (UCI). Km 2 y ½ Autopista La Habana – San Antonio de los Baños, La Habana, Cuba

[2]Universidad de las Ciencias Informáticas (UCI). Km 2 y ½ Autopista La Habana – San Antonio de los Baños, La Habana, Cuba

[*] Autor para correspondencia: dameijeiras@estudiantes.uci.cu

Resumen

Los diversos avances en las ciencias y los grandes volúmenes de datos que se han generado sólo en los últimos años han sobrepasado la capacidad humana para recolectar, almacenar y comprender los mismos sin el uso de las herramientas adecuadas limitando las capacidades de detección de fraudes en las instituciones. Una forma de fraude bancario es el que ocurre con las tarjetas de crédito/débito; estas se han convertido en un método de pago muy popular en las compras online de bienes y servicios. Es por estos motivos que se realizó un análisis de los principales algoritmos de detección de anomalías basados en aprendizaje profundo enfocado en el fraude bancario. Se determinó que las arquitecturas basadas en AEs destacan en tareas no supervisadas y las LSTM para tareas de clasificación.

Palabras clave: anomalía de datos, aprendizaje profundo, fraude bancario, tarjetas de crédito

Abstract

The various advances in science and the large volumes of data that have been generated only in recent years have surpassed the human capacity to collect, store and understand them without the use of the appropriate tools, limiting the fraud detection capabilities of companies. institutions. One form of bank fraud is that which occurs with credit / debit cards; these have become a very popular payment method for online purchases and services. It is for these reasons that an analysis of the main anomaly detection algorithms based on deep learning focused on bank fraud was carried out. Architectures based on AEs were found to excel at unsupervised tasks and LSTMs for classification tasks.

Keywords: *anomaly detection, bank fraud, deep learning, credit cards*

Introducción

Actualmente, estamos viviendo un cambio de paradigma social con el uso y normalización de las tecnologías. Esto ha generado grandes cambios, tanto en particulares, con el aumento del uso de dispositivos y aplicaciones, como en empresas y organizaciones, que se encuentran en un proceso de transformación al nuevo entorno tecnológico, cambiando el modo de trabajo que han tenido hasta ahora. En el nuevo paradigma social, los datos masivos cobran una gran importancia. Se han convertido en información privilegiada para las empresas que buscan sacar el mayor rendimiento. Muchas compañías en sus sistemas de información registran todas las transacciones que se realizan. Si se consideran los diversos avances en las ciencias y los grandes volúmenes de datos que se han generado sólo en los últimos años, es posible notar que estos datos han sobrepasado claramente nuestra capacidad para recolectar, almacenar y comprender los mismos sin el uso de las herramientas adecuadas (Singhal, 2007). Limitando las capacidades de detección de fraudes en las instituciones.

En este contexto es que la Minería de Datos (MD) es una de las soluciones que ayuda a extraer conocimiento a partir de los datos. Este conocimiento puede obtenerse a través de la búsqueda de conceptos, ideas o patrones estadísticamente confiables, que no son evidentes a primera vista, desconocidos anteriormente y que pueden derivarse de los datos originales (Phua et al., 2010), siendo la detección de anomalías una técnica de Minería de Datos con un amplio espectro de aplicaciones de transacciones bancarias y análisis de datos.

Una anomalía es un dato atípico del resto. Esto puede deberse a fallos en mediciones, o a la propia naturaleza del dato. La detección de anomalías (o detección atípica) es la identificación de estos elementos, eventos u observaciones que generan sospechas al diferenciarse significativamente de la mayoría de los datos. Normalmente, estos datos anómalos pueden significar algún tipo de problema o evento raro como: fraude bancario, problemas médicos, defectos estructurales, equipo defectuoso, etc, debido a esto la gran importancia que supone poder identificar a tiempo estos eventos, fundamentales desde una perspectiva empresarial y de salud (Pérez & Mateos, 2018). Esta detección puede ocurrir según (Alcalde, 2018) mediante los siguientes métodos de aprendizaje automáticos según la disponibilidad de los datos:

- métodos supervisados cuando se conoce la existencia de anomalías en los datos, y se sabe cuales son las técnicas usadas son de clasificación supervisada. En este tipo de problemas se tienen dos conjuntos de datos, uno de entrenamiento y otro de test, como se dispone de toda la información, los datos están etiquetados en función de si son anomalía o no, donde se construye un modelo que aprenda a distinguir ente un dato anómalo y uno legítimo.
- métodos semi-supervisados cuando se conoce la existencia de anomalías, pero no se encuentran en el conjunto de datos
- Métodos no supervisados cuando se dispone de anomalías en el conjunto, pero no están etiquetadas, no se conoce a priori si un dato es una anomalía o no, es decir, tanto anomalías como comportamientos legítimos están mezclados. En este campo existen también varias alternativas.

Entre las técnicas de aprendizaje automático, el aprendizaje profundo ha tomado gran popularidad en la comunidad científica, debido a los muy buenos resultados alcanzados en disímiles temas como el procesamiento de imágenes, dígitos, texto y tipos de letras , así como en la detección de anomalías; los enfoques de aprendizaje profundo se utilizan intensamente durante la última década para abordar algunos de los problemas más desafiantes con respecto a la detección de anomalías, como son los relacionados la detección de fraudes bancarios. El aprendizaje profundo es un subconjunto del aprendizaje automático que logra un buen rendimiento y flexibilidad al aprender a representar los datos como una jerarquía anidada de conceptos dentro de las capas de la red neuronal. Las redes neuronales artificiales según (Pal & Shiu, 2004) se encuentran entre las herramientas computacionales que frecuentemente se adoptan para tareas de seguridad por sus ventajas conocidas, como son la adaptabilidad, paralelismo, y aprendizaje;

con ello es posible modificar el peso de las conexiones usando algoritmos de entrenamiento o reglas de aprendizaje y con la actualización de los pesos la red puede optimizar sus conexiones adaptándose a diversos cambios (Petrosino & Maddalena, 2012), lo cual constituye una ventaja sobre los enfoques más tradicionales y menos precisos que proponen los autores (Erich & Zimek, 2011; Hawkins et al., 2002; He et al., 2003; Knorr et al., 2000; Lazarevic & Kumar, 2005; Schölkopf et al., 2001). Las arquitecturas de las redes neuronales ahora van lejos del prototipo biológico, pero todavía producen resultados sobresalientes (Fomin & Bakhshiev, 2019).

Entre distintos campos donde se aplica la detección de anomalía tales como las aplicaciones médicas, la videoseguridad, el procesamiento de imágenes y el monitorio de la red, uno de los más tratados es la detección de fraude, siendo esta según (Chandola et al., 2009), la detección de actividades delictivas que ocurren en organizaciones comerciales como bancos, compañías de tarjetas de crédito, agencias de seguro, compañías de teléfonos celulares y mercado de valores, entre otros.

El fraude es un acto deliberado de engaño para acceder a recursos valiosos (Abdallah et al., 2016). La encuesta mundial sobre delitos económicos de Pricewaterhouse-Coopers (PwC) de 2018 (Lavion & others, 2018 descubrió que la mitad de las 7.200 empresas que encuestaron habían sufrido algún tipo de fraude. El fraude en las telecomunicaciones, las reclamaciones de seguros (de salud, de automóviles, etc.), la banca (reclamaciones de declaraciones de impuestos, transacciones con tarjetas de crédito, etc.) representan problemas importantes tanto en los gobiernos como en las empresas privadas. Detectar y prevenir el fraude no es una tarea sencilla, ya que se trata de un delito adaptativo. Muchos algoritmos tradicionales de aprendizaje automático se han aplicado con éxito en la detección del fraude.

Una forma de fraude bancario es el que ocurre con las tarjetas de crédito/débito; estas se han convertido en un método de pago muy popular en las compras online de bienes y servicios. Este tipo de fraude implica el robo de los datos de una tarjeta de pago y su uso como fuente fraudulenta de fondos en una transacción. El desafío en la detección de fraudes con tarjetas de crédito es que los fraudes no tienen patrones consistentes. El enfoque típico es mantener un perfil de uso para cada usuario y monitorearlos para detectar cualquier desviación. Dado que hay miles de millones de usuarios de tarjetas, esta técnica de perfil de usuario no es muy escalable, es aquí donde la naturaleza del problema permite resolverlo basándose en la detección de anomalías; esta es capaz de detectar patrones de ataques en tiempo

real basado en comportamientos anómalos, que en el caso de los fraudes bancarios podía suponer el robo de identidad o de la tarjeta de crédito. El reto asociado a la detección del fraude es que en la mayoría de los casos requiere una detección y prevención en tiempo real.

Por las razones antes mencionadas, este documento se enfocó en el estudio de los trabajos de detección de anomalías basados en la técnica de aprendizaje profundo, para el tema de detección de fraude bancario. Las contribuciones de este trabajo son una revisión del estado-del-arte de la detección de anomalía y un análisis crítico del mismo. Debido al creciente número de escenarios en los que es posible el fraude.

Materiales y métodos

En esta sección se realiza un análisis del estado-del-arte relacionados a la detección de anomalías, específicamente para la detección de fraude bancario con tarjetas de débito/crédito, que se basan en técnicas de aprendizaje profundo. Para esto se inicia con una explicación de los principales conceptos relacionados con el tema y el funcionamiento de los métodos para la de detección de anomalías.

Detección de Anomalías. Generalidades

La detección de anomalías o *Outliers* se refiere al problema de encontrar datos o grupos de datos que no se ajustan al comportamiento esperado. La detección de *Outliers* tiene uso en una amplia variedad de aplicaciones.

Los *outliers* son los patrones en los datos que no se ajustan a un concepto bien definido de comportamiento normal (Chandola et al., 2007).

En la siguiente Figura 1 se muestra los *Outliers* en un conjunto de datos bi-dimensional. En la figura 1 los datos tienen dos regiones normales: N1 y N2, así como puntos que están lo suficientemente lejos de estas regiones: O1, O2 y O3, estos puntos son los *Outliers*.

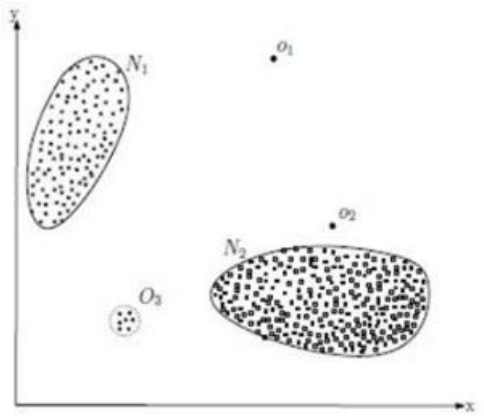

Figura 1. Conjunto de datos bi-dimensional

Un desafío clave en la detección de *Outliers* es que se trata de explorar el espacio invisible. Un acercamiento directo será definir una región que representa un comportamiento normal y declarar cualquier observación en los datos que no pertenece a esta región normal como *Outlier*. Pero varios factores hacen este acercamiento (Cruz Quispe & Rantes García, 2010), al parecer simple, muy desafiante:

- La definición de una región normal que abarque cada comportamiento normal posible es muy difícil
- El límite entre el comportamiento normal y anormal no es a menudo exacto. Así una observación considerada como comportamiento anormal, cerca del límite, puede ser realmente normal y viceversa.
- La noción exacta de un *Outlier* es diferente para diversos dominios del uso. Cada dominio del uso impone un sistema de requisitos y de apremios que dan lugar a una formulación específica del problema para la detección de *Outlier*.
- La disponibilidad de los datos etiquetados para la validación es a menudo un tema importante mientras que desarrolla una técnica de detección del *Outlier*.
- En varios casos, los *Outliers* son el resultado de acciones malintencionadas, los adversarios malévolos se adaptan para hacer que las observaciones anormales aparezcan como normales, de tal modo que la tarea de definir un comportamiento normal se hace más difícil.
- Los datos contienen a menudo ruido que es similar a los *outliers* reales y por lo tanto es difícil de distinguir y quitar.
- Muchas veces el comportamiento normal sigue evolucionando y la noción actual de la conducta normal puede no ser suficientemente representativa en el futuro.

Debido a los desafíos mencionados, el problema de detección *Outliers* en su forma más general, no es fácil de resolver. De hecho, la mayor parte de las técnicas existentes para la detección del mismo simplifican el problema centrándose en una formulación específica. La formulación es inducida por varios factores tales como la naturaleza de los datos, disponibilidad de etiquetado datos, tipo de *Outliers* que se detectarán, etc. A menudo, estos factores son determinados por el dominio del uso en el cual la técnica debe ser aplicada (Cruz Quispe & Rantes García, 2010).

Redes Neuronales Artificiales

El modelo simplificado y abstracto de las redes neuronales artificiales surge de intentar imitar el comportamiento de las neuronas que se encuentran en el cerebro. En las últimas décadas las Redes Neuronales Artificiales (ANN) han recibido un interés particular como una tecnología para minería de datos, puesto que ofrece los medios para modelar de manera efectiva y eficiente problemas grandes y complejos. Los modelos de ANN son dirigidos a partir de los datos, es decir, son capaces de encontrar relaciones (patrones) de forma inductiva por medio de los algoritmos de aprendizaje basado en los datos existentes más que requerir la ayuda de un modelador para especificar la forma funcional y sus interacciones.

Para explicar el funcionamiento de las redes neuronales se comenzará describiendo la red neuronal más simple que existe, una red formada por una única neurona y una única entrada como se muestra en la Figura 2.

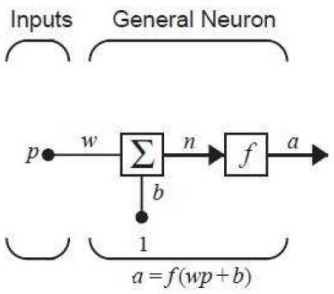

Figura 2. Red Neuronal de única entrada y única salida.

Con este esquema podemos comprender de manera simplificada el funcionamiento de la red neuronal. Los modelos actuales consideran una arquitectura más compleja con múltiples capas y funciones de activación diversas dependiendo del efecto deseado en cada paso del proceso de aprendizaje. En la capa de entrada se considera cada atributo o variable del problema. El proceso de aprendizaje consiste en aprender los pesos de conexión entre dos capas consecutivas lo cual se transfiere de una capa a otra. Para controlar el sobreajuste del modelo se emplean regularizadores apropiados como el conocido dropout. Por 'último, la capa de salida del sistema depende del tipo de función de transferencia que se tenga. Es importante conocer que la función de transferencia la decide el diseñador y que los pesos w y b son ajustables y calculados mediante una regla de aprendizaje conocida como funciones de transferencia. Esta función de transferencia en una red neuronal se elige del problema que se intenta resolver (Durán Suárez, 2017).

Usualmente con una única neurona no será suficiente para resolver la mayoría de problemas prácticos. Para resolver problemas más complejos se tendrá que hacer un uso conjunto de muchas de estas neuronas simples, dando lugar a una verdadera red neuronal, en la que se tendrán cientos o incluso miles de neuronas. Es ahí donde aparece el concepto de capa, que es la agrupación de todas estas neuronas en varios conjuntos dentro de la red neuronal completa (Durán Suárez, 2017), como se muestra en la Figura 3.

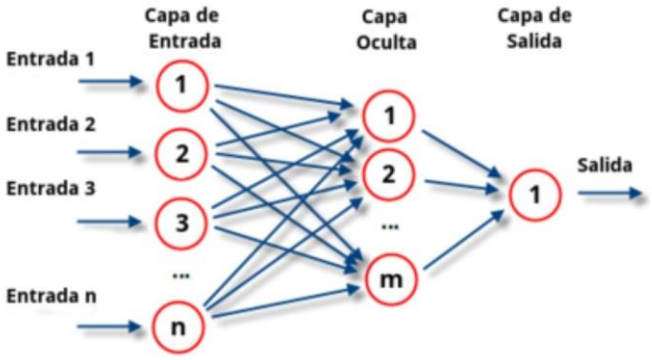

Figura 3. Esquema de red neuronal de tres capas.

El aprendizaje es la clave de la adaptabilidad de la red neuronal y esencialmente es el proceso en el que se adaptan las sinapsis, para que la red responda de un modo distinto a los estímulos del medio. Entrenar una red neuronal es un proceso que modifica los pesos w y el sesgo estadístico b que se obtiene de la interacción entre dos capas, con el fin de que la red pueda a partir de datos de entrada, generar una salida.

Autoenconders

Un autoencoders(AEs) es un modelo computacional basado en redes neuronales compuesto por un codificador y un decodificador, usado para tareas de aprendizaje no supervisado. El principal objetivo de este modelo según es aprender a codificar y decodificar una entrada dada. Esto es particularmente útil cuando se trata de representar los datos con un conjunto más pequeño de características y cuando se trata de reconstruirlos mientras se eliminan las características no deseadas (I. Goodfellow et al., 2016). Existen varios tipos de autocodificadores y pueden emplearse para resolver diversos problemas. Es posible entrenar un autocodificador para que devuelva versiones modificadas de los datos de entrada. Un ejemplo es el autocodificador de eliminación de ruido, que es capaz de aprender a eliminar el ruido de los datos. Los autocodificadores codifican los datos (es decir, los reducen a sus componentes básicos) y luego los reconstruyen mediante el decodificador. Asimismo, existen autocodificadores capaces de generar información (Arroyo & Torres, 2020). Los AEs tienen una estructura simétrica donde las entradas son iguales a la salida como se muestra en la Figura 4.

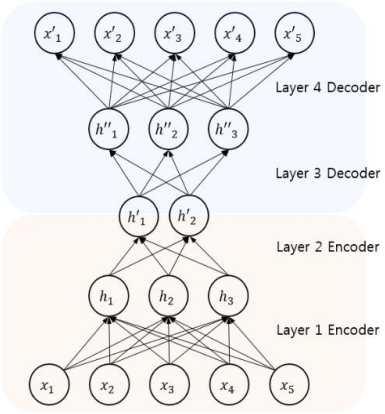

Figura 4. Autoencoders

Las ventajas del uso de AE son la reducción de la dimensionalidad y el aprendizaje de características. Sin embargo, la reducción de las dimensiones y la extracción de características en AE causan algunos inconvenientes. Centrarse en minimizar la pérdida de la relación de los datos en el código de AE provoca la pérdida de alguna relación de datos significativa. Los AEs ofrece una gran oportunidad para construir un detector de fraude incluso en ausencia (o con muy pocos ejemplos) de transacciones fraudulentas. La idea procede del campo más general de la detección de anomalías.

La detección de anomalías (*outliers*) con *Autoencoders* es una estrategia no supervisada para identificar anomalías cuando los datos no están etiquetados, es decir, no se conoce la clasificación real (anomalía - no anomalía) de las observaciones. Si bien esta estrategia hace uso de *Autoencoders*, no utiliza directamente su resultado como forma de detectar anomalías, sino que emplea el error de reconstrucción producido al revertir la reducción de dimensionalidad. El error de reconstrucción como estrategia para detectar anomalías se basa en la siguiente idea: los métodos de reducción de dimensionalidad permiten proyectar las observaciones en un espacio de menor dimensión que el espacio original, a la vez que tratan de conservar la mayor información posible. La forma en que consiguen minimizar la perdida global de información es buscando un nuevo espacio en el que la mayoría de observaciones puedan ser bien representadas (Rodrigo, 2020).

El método de *Autoencoders* crean una función que mapea la posición que ocupa cada observación en el espacio original con el que ocupa en el nuevo espacio generado. Este mapeo funciona en ambas direcciones, por lo que también se puede ir desde el nuevo espacio al espacio original. Solo aquellas observaciones que hayan sido bien proyectadas podrán volver a la posición que ocupaban en el espacio original con una precisión elevada (Rodrigo, 2020).

Dado que la búsqueda de ese nuevo espacio ha sido guiada por la mayoría de las observaciones, serán las observaciones más próximas al promedio las que mejor puedan ser proyectadas y en consecuencia mejor reconstruidas. Las observaciones anómalas, por el contrario, serán mal proyectadas y su reconstrucción será peor. Es este error de reconstrucción (elevado al cuadrado) el que puede emplearse para identificar anomalías (Rodrigo, 2020).

Los AEs han sido de gran utilidad para la detección de fraude bancario con tarjetas de débito/crédito, al ser esta una tarea de aprendizaje no supervisado y unos de los referentes en los últimos años (Pumsirirat & Yan, 2018; Zou et al., 2019). (Paula et al., 2016) utilizaron AE para implementar la detección de en fraude financiero y el lavado de dinero para las empresas brasileñas en las reclamaciones de impuestos a la exportación. Los autores de (Pumsirirat & Yan, 2018) propusieron un método para la detección de fraude en tarjetas de crédito que consiste en clasificar una petición de transferencia bancaria en tiempo real usando un AE, el cual se entrena teniendo en cuenta la información de transacciones realizadas con anterioridad. (P. Zheng et al., 2019; Y.-J. Zheng et al., 2018) proponen arquitecturas de red donde se combinan estructuras dististas de redes neuronales utilizando en ambos casos AEs para la extracción de características una red de tipo GAN como clasificador, detectando actividad en transacciones y usuarios fraudulentos. (Zou et al., 2019) utiliza una arquitectura basada en AE con un algoritmo de sobre muestreo para sintetizar nuevas muestras en una clase minoritaria, obteniendo mejoras en la precisión de la clasificación de la clase minoritaria de los conjuntos de datos desequilibrados.

Restricted Boltzmann Machina **(RBM)**

La RBM es un tipo diferente de modelo de RNA que puede aprender la distribución de probabilidad del conjunto de entrada (Qiu et al., 2014). Los RBM se utilizan principalmente para la reducción de la dimensionalidad, la clasificación y el aprendizaje de características. El RBM es un modelo gráfico bipartito no dirigido que consta de dos capas: la visible y la oculta. Las unidades de la capa no están conectadas entre sí. Cada celda es un punto computacional que procesa la entrada. Cada unidad toma decisiones estocásticas sobre la transmisión de los datos de entrada o no. Las entradas se multiplican por pesos específicos, se añaden ciertos valores de umbral (sesgo) a los valores de entrada y, a continuación, los valores calculados pasan por una función de activación. En la etapa de reconstrucción, los resultados de las salidas vuelven a entrar en la red como la entrada, y luego salen de la capa visible como la salida. Se comparan los valores de la entrada anterior y los valores después de los procesos como se muestra en la Figura 5. El objetivo de la comparación es reducir la diferencia. El aprendizaje se realiza varias veces en la red (Qiu et al., 2014). La RBM es un modelo gráfico bicapa, bipartito y no dirigido que consta de dos capas: la visible y la oculta. Las capas no están conectadas entre sí. La desventaja de los RBM es su difícil entrenamiento.

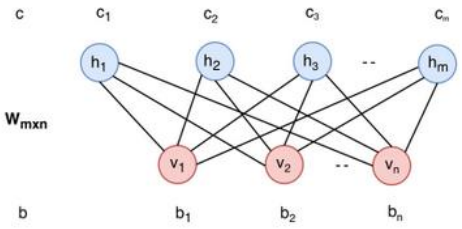

Figura 5. *Restricted Boltzmann Machina*

El método propuesto en (Pumsirirat & Yan, 2018), utiliza una RBM para la clasificación de fraude en tarjetas de crédito. En este método se valida una petición de transferencia bancaria en tiempo real entrenando la RBM teniendo en cuenta la información de transacciones realizadas con anterioridad, donde se obtuvo resultados satisfactorios en conjuntos de datos de aprendizaje supervisado o para una base de datos de historial de detección de fraudes con tarjetas de crédito. De esta forma se puede identificar si una transacción es fraudulenta o legítima si contrasta contra un patrón de comportamiento.

Redes Neuronales Convolucionales

Una red neuronal convolucional según (Durán Suárez, 2017) es un tipo de red multicapa que consta de diversas capas convolucionales y de pooling (submuestreo) alternadas, y al final tiene una serie de capas full-connected como una red de tipo perceptrón multicapa como se puede ver en la Figura 6.

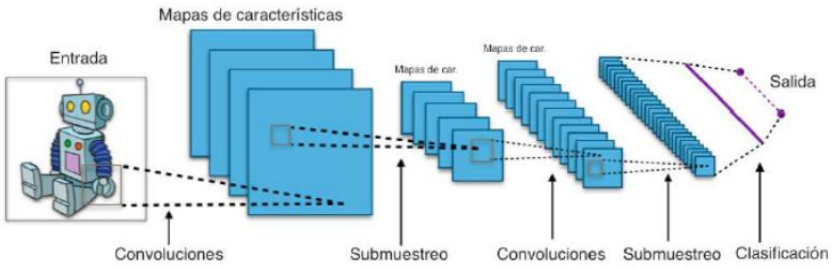

Figura 6. Arquitectura de una Red Neuronal Convolucional

Las capas convolucionales tienen filtros (o kernels) cuyas dimensiones son $l \times p \times q$. Las configuraciones de las dimensiones del kernel son elegidas por el diseñador (generalmente q suele ser igual a r). Cada filtro se aplica mediante una operación de convolución con lo cual se obtiene un mapa o tensor de rasgos o características de tamaño $(m - l + 1) \times (n - p + 1) \times q$. Después cada mapa es sub-muestreado en la capa de pooling con las operaciones estadísticas de máximo, mínimo o media sobre regiones contiguas de un tamaño dado $k \times k$. Antes o después del submuestreo, se aplica una función de activación sigmoidal más un sesgo para cada mapa de rasgos.

Las redes neuronales convolucionales (CNN), son la elección popular de las redes neuronales para el análisis de imágenes visuales, aunque su capacidad permite extraer características ocultas en datos de alta dimensión con estructura compleja y ha permitido su uso como extractores de características en la detección de valores atípicos para el conjunto de datos secuenciales y de imágenes. Esta más enfocado a tareas de aprendizaje supervisado sobre conjuntos de datos previamente etiquetados (Abroyan, 2017; Fu et al., 2016; P. Zheng et al., 2019).

Modelos Generativos

El objetivo principal de las *Redes generativas antagónicas*(GAN) (I. J. Goodfellow et al., 2014) es generar datos desde cero. Para ello las GAN emplean dos redes neuronales y las enfrentan mutuamente. La primera red es el "generador" y la segunda es el "discriminador". Ambas redes fueron entrenadas con un mismo conjunto de datos, pero la primera debe intentar crear variaciones de los datos que ya ha visto, en el caso de los rostros de personas que no existen, debe crear variaciones de los rostros que ya ha visto. La red discriminatoria debe identificar si ese rostro que está viendo forma parte del entrenamiento original o si es un rostro falso que creó la red generativa. Mientras más lo hace, la red generativa se hace mejor creando y a la red discriminadora se le hace más difícil detectar si el rostro es falso. La red generadora necesita la discriminadora para saber cómo crear una imitación tan realista que la segunda no logre distinguir de una imagen real. Una red generadora por sí sola crearía solo ruido aleatorio, el concepto es que la red discriminadora hace de guía sobre cuáles imágenes crear y ayuda a la red generativa a aprender los aspectos que comprenden una imagen real. El modelo entrena a ambas redes y las enfrenta en una dura competición para que se mejoren a sí mismas. Eventualmente, el discriminador será capaz de identificar la más pequeña diferencia entre lo que es real y lo que fue generado, y la red generativa será capaz de crear imágenes que el discriminador no puede distinguir Figura 7.

Los modelos generativos pretenden aprender la distribución exacta de los datos para generar nuevos puntos de datos con algunas variaciones. Una variante de la arquitectura GAN, conocida como autocodificadores antagónicos (AAE) (Makhzani et al., 2015), que utiliza el entrenamiento adversarial para imponer una prioridad arbitraria en el código latente aprendido dentro de las capas ocultas del autocodificador, también ha demostrado aprender la distribución de entrada de forma eficaz. Aprovechando esta capacidad de aprendizaje de las distribuciones de entrada, varios marcos de detección de anomalías basados en redes antagónicas generativas (GAN-AD) (D. Li et al., 2018; P. Zheng et al., 2019) propuestos han demostrado ser eficaces en la identificación de anomalías en conjuntos de datos altamente dimensionales y complejos.

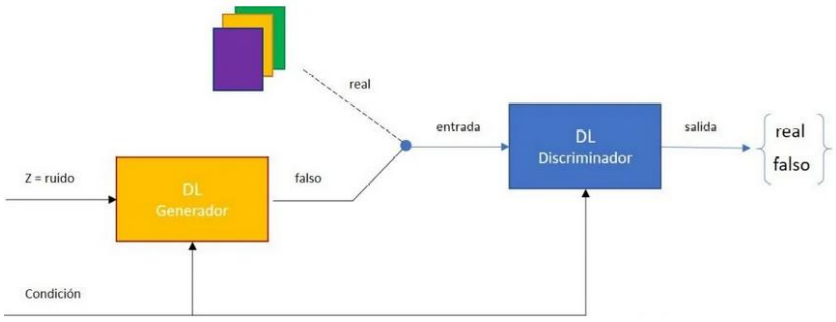

Figura 7. Arquitectura de una red GAN

Modelos Secuenciales

En la literatura, las Redes Neuronales Recurrentes (RNN) se han utilizado sobre todo en datos secuenciales, como datos de series temporales, datos de audio y habla, y lenguaje. A diferencia de las redes de avance, las RNN utilizan una memoria interna para procesar las entradas entrantes. Las RNN se utilizan en el análisis de datos de series temporales en varios campos. En general, la RNN procesa las series de entrada de una en una, durante su funcionamiento. Las unidades de la capa oculta contienen información sobre la historia de la entrada en el vector de estado. Las RNN se pueden entrenar utilizando el método de propagación en a través del tiempo (BPTT). Con el método BPTT, la diferenciación de la pérdida en cualquier momento refleja los pesos de la red en el paso de tiempo anterior. El entrenamiento de las RNN es más difícil que el de las redes neuronales de avance (FFNN) y el periodo de entrenamiento de las RNN es más largo (Ozbayoglu et al., 2020). Las redes neuronales recurrentes (RNN) son

arquitecturas de aprendizaje profundo cíclicas que abordan secuencias arbitrarias de patrones de entrada. Las arquitecturas RNN incluyen la de memoria larga a corto plazo (LSTM) (Hochreiter, 1997)y la unidad recurrente cerrada (Cho et al., 2014).

La red LSTM es un tipo diferente específicamente pensada para el análisis de datos secuenciales. La ventaja de las redes LSTM reside en el hecho de que tanto los valores a corto como a largo plazo de la red pueden ser recordados. La unidad LSTM se compone de células con puertas de entrada, salida y olvido. Estas tres puertas regulan el flujo de información. Con estas características, cada célula recuerda los valores deseados durante intervalos de tiempo arbitrarios (Ozbayoglu et al., 2020). En cambio, las Redes Neuronales recurrentes GRU están diseñadas similarmente a las LSTM, con la diferencia que, para solucionar el problema de gradiente de fuga, GRU utiliza dos puertas (actualización y reinicio). Los funcionamientos de las puertas son parecidos a las LSTM. Básicamente, estas dos puertas deciden qué información se debe pasar a la salida. Lo especial de ellos es que pueden ser entrenados para mantener la información a largo plazo.

Las redes neuronales de este tipo se han utilizado tanto para clasificación multiclase como para una sola clase. La técnica básica para detectar anomalías haciendo uso de clasificadores multiclase mediante redes neuronales se fundamenta en dos fases. Primero se entrena con los datos destinados a ello para aprender las diferentes clases. Tras eso, cada instancia de prueba se pasa como una entrada a la red neuronal. Si la red acepta la entrada de prueba, es normal y si la red la rechaza, es una anomalía. Se construye una red neuronal multicapa, que tendrá el mismo número de neuronas de entrada y de salida. En la fase de entrenamiento se comprimirán los datos en las capas ocultas y durante la fase de prueba, se reconstruirá cada instancia de los datos haciendo uso de la red entrenada para obtener la salida (García Costa, 2018).

En el tema de detección de anomalías para fraude bancario (X. Li et al., 2018) trabajan las transacciones como subsecuencias de eventos por cuenta. Los autores separan las cuentas en conjuntos según el número de transacciones y entrenan una GRU independiente para cada conjunto. Una de las limitaciones de este enfoque es que, en un escenario de producción, las secuencias son ilimitadas y, por tanto, no se puede asignar las cuentas a los conjuntos de forma automática. En particular, no está claro qué ocurrirá con los estados ocultos cuando una cuenta cambie de conjunto. (Jurgovsky et al., 2018) emplean una arquitectura basada en LSTM para realizar la clasificación de

secuencias para la detección de fraudes en entornos de comercio electrónico y cara a cara, integrando también métricas de agregación de características, es decir, perfiles. (Branco et al., 2020) para aprovechar mejor la información definen instancias puntuables y no puntuables en el entrenamiento considerando secuencias cortas y largas de 5 y 10 pasos de tiempo. (Alghofaili et al., 2020) propone un método con una LSTM de base donde se utiliza un conjunto de datos reales de fraudes con tarjetas de crédito y los resultados se comparan con un modelo de AE, donde se obtienen resultados de 99,95% de precisión en menos de un minuto para el conjunto de datos.

Conclusiones

Una de las formas de fraude bancario es el que ocurre con las tarjetas de crédito/débito; estas se han convertido en un método de pago muy popular en las compras online de bienes y servicios. El enfoque típico en la detección de fraudes con tarjetas es mantener un perfil de uso para cada usuario y monitorearlos para detectar cualquier desviación, lo que supone un costo muy elevado dado que hay miles de millones de usuarios de tarjetas, siendo esta técnica poco escalable, ahí es donde el uso del aprendizaje profundo ha ganado cada vez más importancia; por su eficacia y precisión garantizan una mayor seguridad. Esto ha permitido la inclusión de este tipo de aprendizaje en la mayoría de las aplicaciones de reconocimiento de patrones y tareas de minería de datos, con resultados significativo en la mayoría de los casos en comparación con los métodos tradicionales.

El problema de detección de anomalías en transacciones con tarjetas de crédito/débito se ha abordado con distintas arquitecturas de redes, destacando el uso de AEs, con su estructura básica o en combinación con otros tipos de redes profundas, esto gracias a sus garantías en tareas de aprendizaje no supervisado, que permite una detección de datos atípicos en tiempo real, sin tener un conjunto previamente etiquetado.

Las otras arquitecturas de redes también presentan resultados satisfactorios en tareas de detección de anomalías de datos de forma general, ya sea como estructura base o combinación con otra. En el caso de fraude bancario con tarjeta de crédito el enfoque es en encontrar infracciones sobre base de datos previamente categorizadas, recientemente resaltan los resultados de los modelos basados en LSTM.

El aprendizaje profundo es la solución para la detección de fraudes bancarios con tarjetas de crédito/débito como alternativa a la cantidad de datos que se generan constantemente las transacciones bancarias, sustituyendo los altos costos que implicarían los métodos tradicionales. La arquitectura seleccionada como propuesta de solución a un determinado problema debe estar de acorde a la naturaleza de la tarea solicitada y a las condiciones del conjunto de datos.

El reto más importante que deben asumir los desarrolladores de modelos de prevención de fraude bancario basado en aprendizaje profundo es el de construir modelos predictivos cada vez más flexibles y adaptables a condiciones cambiantes con rapidez. Que tengan la capacidad de detectar eventos únicos o emergentes es el objetivo a futuro.

Referencias

Abroyan, N. (2017). Neural networks for financial market risk classification. *Frontiers in Signal Processing*, *1*(2).

Alcalde, A. (2018). *Aprendizaje no Supervisado y Detección de Anomalías: ¿Qué es una Anomalía?* El Baúl del Programador. https://elbauldelprogramador.com/aprendizaje-nosupervisado-anomalias/

Alghofaili, Y., Albattah, A., & Rassam, M. A. (2020). A Financial Fraud Detection Model Based on LSTM Deep Learning Technique. *Journal of Applied Security Research*, *15*(4), 498-516.

Arroyo, A. C., & Torres, G. S. (2020). Remoción de lluvia en imágenes por medio de una arquitectura de autoencoder. *Investigación e Innovación en Ingenierías*, *8*(1), 140-167.

Branco, B., Abreu, P., Gomes, A. S., Almeida, M. S., Ascensão, J. T., & Bizarro, P. (2020). Interleaved Sequence RNNs for Fraud Detection. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3101-3109.

Chandola, V., Banerjee, A., & Kumar, V. (2007). Outlier detection: A survey. *ACM Computing Surveys*, *14*, 15.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, *41*(3), 1-58.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Cruz Quispe, L. M., & Rantes García, M. T. (2010). *Detección de fraudes usando técnicas de clustering*.

Durán Suárez, J. (2017). *Redes neuronales convolucionales en R: Reconocimiento de caracteres escritos a mano*. Universidad de Sevilla.

Erich, H.-P. K. P. K., & Zimek, S. A. (2011). Interpreting and unifying outlier scores. *11th SIAM International Conference on Data Mining (SDM), Mesa, AZ*.

Fomin, I. S., & Bakhshiev, A. V. (2019). Springer International Publishing,. En *Research on Convolutional Neural Network for Object Classification in Outdoor Video Surveillance System, en Cham* (pp. 221-229).

Fu, K., Cheng, D., Tu, Y., & Zhang, L. (2016). Credit card fraud detection using convolutional neural networks. *International Conference on Neural Information Processing*, 483-490.

García Costa, H. (2018). *Herramienta para la detección de anomalías en el suministro eléctrico en el ámbito asistencial* [PhD Thesis].

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1). MIT press Cambridge.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks. *arXiv preprint arXiv:1406.2661*.

Hawkins, S., He, H., Williams, G., & Baxter, R. (2002). Outlier detection using replicator neural networks. *International Conference on Data Warehousing and Knowledge Discovery*, 170-180.

He, Z., Xu, X., & Deng, S. (2003). Discovering cluster-based local outliers. *Pattern Recognition Letters*, *24*(9-10), 1641-1650.

Hochreiter, S. (1997). JA1 4 rgen Schmidhuber (1997).“Long Short-Term Memory”. *Neural Computation*, *9*(8).

Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P.-E., He-Guelton, L., & Caelen, O. (2018). Sequence classification for credit-card fraud detection. *Expert Systems with Applications*, *100*, 234-245.

Knorr, E. M., Ng, R. T., & Tucakov, V. (2000). Distance-based outliers: Algorithms and applications. *The VLDB Journal*, *8*(3-4), 237-253.

Lazarevic, A., & Kumar, V. (2005). Feature bagging for outlier detection. *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 157-166.

Li, D., Chen, D., Goh, J., & Ng, S. (2018). Anomaly detection with generative adversarial networks for multivariate time series. *arXiv preprint arXiv:1809.04758*.

Li, X., Yu, W., Luwang, T., Zheng, J., Qiu, X., Zhao, J., Xia, L., & Li, Y. (2018). Transaction fraud detection using gru-centered sandwich-structured model. *2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design ((CSCWD))*, 467-472.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., & Frey, B. (2015). Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*.

Ozbayoglu, A. M., Gudelek, M. U., & Sezer, O. B. (2020). Deep learning for financial applications: A survey. *Applied Soft Computing*, 106384.

Pal, S. K., & Shiu, S. C. K. (2004). Foundations of Case-Based Reasoning. *JohnWiley & Sons, Inc*.

Paula, E. L., Ladeira, M., Carvalho, R. N., & Marzagao, T. (2016). Deep learning anomaly detection as support fraud investigation in brazilian exports and anti-money laundering. *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 954-960.

PÉREZ, E. R., & MATEOS, A. (2018). *Análisis y Detección de Fraude Fiscal Mediante Técnicas de Aprendizaje Automático*.

Petrosino, A., & Maddalena, L. (2012). *Neural Networks in Video Surveillance: A Perspective View*. https://doi.org/10.1201/b11631-4

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*.

Pumsirirat, A., & Yan, L. (2018). Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine. *International Journal of advanced computer science and applications*, *9*(1), 18-25.

Qiu, X., Zhang, L., Ren, Y., Suganthan, P. N., & Amaratunga, G. (2014). Ensemble deep learning for regression and time series forecasting. *2014 IEEE symposium on computational intelligence in ensemble learning (CIEL)*, 1-6.

Rodrigo, J. (2020). *Detección de anomalías: Autoencoders y PCA*. Cienciadedatos.net. https://www.cienciadedatos.net/documentos/52_deteccion_anomalias_autoencoder_pca.html

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural computation*, *13*(7), 1443-1471.

Singhal, A. (2007). Data modeling and data warehousing techniques to improve intrusion detection. En *Data warehousing and data mining techniques for cyber security* (pp. 69-82). Springer.

Zheng, P., Yuan, S., Wu, X., Li, J., & Lu, A. (2019). One-class adversarial nets for fraud detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 1286-1293.

Zheng, Y.-J., Zhou, X.-H., Sheng, W.-G., Xue, Y., & Chen, S.-Y. (2018). Generative adversarial network based telecom fraud detection at the receiving bank. *Neural Networks, 102*, 78-86.

Zou, J., Zhang, J., & Jiang, P. (2019). Credit card fraud detection using autoencoder neural network. *arXiv preprint arXiv:1908.11553*.