Universidad del Valle de Guatemala
Facultad de ingeniería

Security Data Science
Proyecto 2

Entrenamiento Incremental en Modelos de Deep Learning y Machine Learning

Andrés de la Roca

Guatemala, 2024

Investigación Teórica

Típicamente el entrenamiento de los modelos de machine learning toma un enfoque de entrenamiento "total", es decir que el modelo desarrollado se entrena sobre todo un dataset en una sola sesión. Este acercamiento es útil en ocasiones donde el conjunto de datos es pequeño y puede ser procesado de manera rápida, sin embargo, esta opción no es óptima para conjuntos de datos donde se tiene una cantidad grande datos como para ser procesados en una sola sesión o cuando los datos son cambiantes, como por ejemplo con datos de redes sociales o datos del mercado de la bolsa.
El entrenamiento incremental busca solventar este problema, creando oportunidades para que el modelo pueda aprender y mejorar progresivamente, esto se hace mediante la entrada de nuevos datos, es decir que para datos que se van actualizando constantemente se va a poder observar una adaptación correcta del modelo a estos nuevos datos (Ade, Deshmukh., 2013).

Este enfoque también tiene sus limitaciones y vulnerabilidades a tomar en cuenta, entre ellas, el "olvido catastrófico", en el cual el modelo presenta una deficiencia en la retención de datos pasados al entrenarse con datos nuevos, lo cual puede presentarse en forma de una disminución de la precisión general del modelo. También, este enfoque puede llevar al modelo a tener una tendencia hacia el sobreajuste, ya que podría llegar cierto momento durante su entrenamiento en el que el modelo podría sobre ajustar sus parámetros basado en un cambio abrupto de datos que podrían llegar a no representar la distribución normal de los datos.

Vale la pena mencionar, dentro del mismo enfoque del entrenamiento incremental se pueden tomar ciertas variaciones según el contexto del problema a resolver. (Van de Ven, Tuytelaars, Tolias., 2022)

Con todo lo anteriormente expuesto, se puede observar que el entrenamiento incremental es una gran oportunidad para el desarrollo de la inteligencia artificial, abre la posibilidad del mejoramiento continuo de modelos existentes para poder alcanzar una precisión extremadamente alta y ganando experiencia con cada interacción que el modelo tiene con nuevos datos.

Descripción de implementación

Para el entrenamiento del modelo se dividió el conjunto de datos en 70% como conjunto de entrenamiento y 30% como conjunto de pruebas. El conjunto en general se dividirá en transacciones del año 2019 y año 2020, de esta manera el primer entrenamiento se realizará con transacciones del año 2019 y luego el segundo entrenamiento de manera incremental se realizará con las del año 2020.

Vale la pena mencionar que el desbalance extremo en el conjunto de datos entre transacciones legítimas y fraudulentas es de 99.47% a 0.52% respectivamente, para lograr disminuir el impacto negativo que esto podría llegar a tener se utilizó la técnica de sobre muestreo SMOTE (Synthetic Minority Over-sampling Technique), el cual genera nuevas muestras sintéticas de la clase minoritaria, es decir de las transacciones fraudulentas.

El modelo incremental de random forest utilizó los siguientes parámetros para su entrenamiento:

El modelo utiliza inicialmente 10 árboles de decisión para el entrenamiento inicial, luego en el segundo entrenamiento se le agregan 10 más para un total de 20 árboles de decisión en el modelo.
La profundidad máxima de cada árbol se limitó a 2 y el número mínimo de muestras requeridas para dividir un nodo interno se estableció en 5.

El parámetro del modo calentamiento se establece como verdadero, este es clave para poder entrenar de manera incremental el modelo.

El modelo incremental de SVM utilizó los siguientes parámetros para su entrenamiento:

Como clasificador se utiliza SGD, el cual permite hacer entrenamiento incremental mientras se mantienen las características claves de un SVM normal, se utilizó el parámetro de función de pérdida "hinge" debido a que este es un problema de clasificación binario.
Se aplica regularización L2 para evitar el sobreajuste por medio del parámetro de penalización.
En este caso el propio clasificador tiene habilitado la habilidad de hacer entrenamiento incremental simplemente utilizando la función partial_fit, la cual hace un entrenamiento incremental con datos nuevos, utilizando lo previamente aprendido por el modelo, sin necesidad de especificar el inicio del modo de calentamiento.

Análisis de resultados

Para el modelo que utiliza Random Forest se obtuvieron los siguientes resultados:

- ROC AUC Score:
    - Sobre los datos de 2019, el área bajo la curva es de 88.88%, lo que indica que el modelo tiene capacidad alta para distinguir entre las transacciones fraudulentas y legítimas. En 2020, este puntaje aumentó ligeramente a 89.62%, lo que sugiere una mejora en la capacidad de clasificación del modelo después del entrenamiento incremental. Sin embargo, este incremento es mínimo, por lo que quizás para futuras iteraciones se debería de considerar alimentarlo con más datos para mejorar este aspecto del modelo.
- Recall Score:
    - En 2019, el recall es del 84.28%, esta métrica representa la tasa de precisión en la que el modelo identifica correctamente las transferencias ilegítimas. Luego del entrenamiento incremental, utilizando los datos del 2020, se observó una ligera disminución en el puntaje con 83.89%, sin embargo, este puntaje sigue siendo bastante alto, por lo que sugiere que el modelo sigue siendo bastante efectivo en la identificación de las transferencias ilegítimas. Como posible punto de mejora en este aspecto, de manera similar al AUC Score, se podría considerar utilizar una mayor cantidad de datos o hasta incluso una división más granular en los entrenamientos incrementales, es decir en lugar de realizarlo por medio de incrementos anuales, se podrían realizar de manera semestral o trimestral para observar alguna mejora
- F1 Score:
    - El puntaje F1 que es una combinación de ambas métricas anteriores nos muestra que en general hay una ligera mejora en el rendimiento del modelo de 88.34% a 88.99% con la data de 2019 y 2020 respectivamente, lo que demuestra que si se puede lograr una mejora general del rendimiento del modelo utilizando el método de entrenamiento incremental.

Para el modelo que utiliza SVM se obtuvieron los siguientes resultados:

- ROC AUC Score:

    - En 2019, el área bajo la curva es de 89.05%, lo que muestra que el modelo tiene una alta capacidad para la distinción de ambos casos de transacciones. En 2020, se observa una mejora marginal, 89.39% después del entrenamiento incremental. Esto sugiere que aún hay margen de mejora con este modelo, quizás empleando otras características en el conjunto de datos para que el modelo pueda más fácilmente realizar la distinción de clases de transacciones.
- Recall Score:
    - Para el primer entrenamiento con los datos de 2019, el recall es del 83.16%, lo que indica que el modelo identificó correctamente el 83.16% de las transferencias fraudulentas. De manera similar al AUC Score, el recall aumentó ligeramente al 83.55% tras el entrenamiento incremental, por lo que aún hay margen de mejora para este aspecto.
- F1 Score:
    - En el puntaje general del modelo, se refleja lo anteriormente visto en los puntajes anteriores, tras el entrenamiento incremental se observa una mejora marginal en el modelo, 88.37% y 88.73% para 2019 y 2020 respectivamente.

A través de los resultados de ambos modelos podemos observar que cada uno pasó por una ligera mejora tras realizar el entrenamiento incremental, sin embargo, valdría la pena en futuras iteraciones observar cómo se comporta con diferentes variables en el conjunto de datos, con diferentes algoritmos y metodologías de machine learning e incluso con la inclusión de más datos, para observar si estas mejoras marginales observadas en ambos modelos puede aumentar para poder crear un modelo aún más robusto.

Metodología propuesta

Para hacer una elección entre metodologías de entrenamiento se ideó la siguiente metodología la cual evalúa algunos de los aspectos más importantes a considerar para elegir entre reentrenamiento total o entrenamiento incremental tomando en cuenta sus ventajas y desventajas.

1. Naturaleza de las transacciones fraudulentas
   Para examinar la naturaleza de estas transacciones nos tenemos que preguntar: ¿Los patrones de fraude cambian con el tiempo o permanecen relativamente estables? ¿Hay nuevos tipos de fraudes que surgen con frecuencia? Por lo general, mientras más cambiante son los datos la opciona óptima a elegir sería un entrenamiento incremental, ya que esto le podría permitir a los modelos adaptarse de manera correcta a los cambios, por otro lado, si no se tuvieran datos estables a través del tiempo, se debería de considerar utilizar un entrenamiento total.

2. Disponibilidad de datos
   Determinar la disponibilidad y frecuencia de actualización de los datos es importante. Si tiene un acceso regular a datos nuevos de manera constante, el reentrenamiento incremental puede ser adecuado para mantener al modelo actualizado y que se adapte a nuevas tendencias dentro de los datos. Por otro lado, si los datos nuevos son escasos y la distribución de datos no cambia significativamente con el tiempo, un reentrenamiento total puede resultar más apropiado. Es decir, si se logra tener acceso constante en periodos de tiempo frecuentes a datos sobre transacciones podría resultar más apropiado el entrenamiento incremental y podrá otorgar mejores resultados. Por otro lado, si se tuviera un acceso más limitado a los datos de

transacciones y estos no se actualizarán de manera constante, el reentrenamiento total podría ser la mejor elección.

3. Evaluación de recursos computacionales:
   Hay que considerar la capacidad computacional necesaria para reentrenar el modelo, ya que el reentrenamiento total puede requerir muchos más recursos, ya que implica utilizar todo el conjunto de datos para volver a entrenar el modelo. Por otro lado, el reentrenamiento incremental puede ser más eficiente en términos de tiempo y recursos debido a su naturaleza de entrenamientos más cortos y con conjuntos de datos más reducidos. Vale la pena mencionar que, en caso no se tenga un alto volumen de datos, un reentrenamiento total también podría ser una opción válida, debido a que no consumirá tanto tiempo y recursos.

4. Comparación de métricas de rendimiento a lo largo del tiempo:
   Realizar un seguimiento de las métricas de rendimiento del modelo a lo largo del tiempo utilizando datos pasados y datos actualizados podría mostrar si un reentrenamiento total es necesario o si la metodología de entrenamientos incrementales está funcionando correctamente. Por ejemplo, si con el paso del tiempo con datos posteriores a 2020, el modelo mantuviera un rendimiento estable o mostrará mejoras, esto demostraría que el entrenamiento incremental está siendo funcional y que es válido mantener esta metodología. Sin embargo, en el caso contrario, en donde las métricas muestran una disminución en la precisión en la detección de transacciones fraudulentas se podrá determinar que un reentrenamiento total es más adecuado.

Con esta metodología se puede determinar si podría resultar favorable realizar un tipo de entrenamiento o el otro dependiendo de la situación y contexto en el que se encuentre en el proyecto al momento de realizar estas evaluaciones sobre el modelo y sus métricas.

Conclusiones

- Se deben de utilizar las estrategias adecuadas de entrenamiento según el contexto en el que se encuentre el proyecto y sus necesidades correspondientes.
- Se deben de manejar de manera adecuada los datos desequilibrados para no afectar el entrenamiento y poder mantener un buen rendimiento durante la evaluación de datos reales.
- La ingeniería de características es un aspecto importante a tener en cuenta cuando se desea obtener una mejor correlación entre las diferentes variables dentro de un conjunto de datos y por lo tanto mejorar el rendimiento general del modelo.
- El entrenamiento incremental es una metodología muy importante y util cuando se tienen datos con una alta variabilidad y poca estabilidad a lo largo del tiempo.

Recomendaciones

Algunas de las recomendaciones para futuras iteraciones tomando en cuenta lo visto durante el desarrollo del proyecto pueden ser:
En primer lugar, contar con conjuntos de datos que amplíen el plazo de tiempo en el que se tomaron las transacciones, es decir que tengas de años pasados y posteriores para de esta manera poder reflejar una diversidad mayor en las transacciones.

Asimismo, la selección de nuevas características dentro del conjunto de datos existente, para de esta manera poder garantizar que el modelo pueda relacionarse de mejor manera con los datos y poder obtener un mejor rendimiento en sus predicciones.
Otra estrategia que podría resultar clave para el mejoramiento del proyecto en general es el uso de otros modelos como Naive Bayes o redes neuronales para observar cómo se comportan y tener un mayor panorama en cuanto a cuál podría ser la mejor opción si en algún momento se desea ampliar el alcance del proyecto.

Referencias bibliográficas

Awan, A., (2023) What is Incremental Learning?
https://www.datacamp.com/blog/what-is-incremental-learning

Shilton, A., Palaniswami, M., Ralph, D., & Tsoi, A. C. (2005). Incremental training of support vector machines. IEEE transactions on neural networks, 16(1), 114-131.

Domeniconi, C., & Gunopulos, D. (2001, November). Incremental support vector machine construction. In Proceedings 2001 ieee international conference on data mining (pp. 589-592). IEEE.

Hu, C., Chen, Y., Hu, L., & Peng, X. (2018). A novel random forests based class incremental learning method for activity recognition. Pattern Recognition, 78, 277-290.

Van de Ven, G. M., Tuytelaars, T., & Tolias, A. S. (2022). Three types of incremental learning. Nature Machine Intelligence, 4(12), 1185-1197.