UNIVERSIDAD DEL VALLE DE GUATEMALA

Facultad de Ingeniería

Investigación sobre Entrenamiento incremental en los algoritmos para proyecto final Security Data Science

Marco Antonio Jurado Velasquez
Carné 20308

Guatemala
2024

Contenido

Introducción

En el ámbito del aprendizaje automático, la adaptación eficiente de modelos predictivos ante la llegada continua de nuevos datos es crucial. El entrenamiento incremental permite actualizar los modelos sin reentrenarlos desde cero, optimizando recursos computacionales, reduciendo el tiempo de procesamiento y mejorando la capacidad de adaptación a cambios en los patrones de datos.

Tradicionalmente, los algoritmos de aprendizaje automático requieren todo el conjunto de datos durante el entrenamiento, lo cual es impracticable con grandes volúmenes de datos o en aplicaciones en tiempo real. El entrenamiento incremental emerge como una solución ante esta limitación, aplicándose en algoritmos robustos como Redes Neuronales Artificiales (ANN), LightGBM, XGBoost, Random Forest y Máquinas de Vectores de Soporte (SVM).

Este trabajo investiga la viabilidad y eficacia del entrenamiento incremental en estos modelos, utilizando un conjunto de datos de transacciones de tarjetas de crédito para clasificar actividades normales y fraudulentas. Se destacan las ventajas prácticas del entrenamiento incremental y se examinan los desafíos técnicos, como la estabilidad del modelo y la prevención del olvido catastrófico, donde un modelo pierde información previamente aprendida al adaptarse a nuevos datos.

A través de una metodología que incluye revisión de literatura, implementación práctica y evaluación rigurosa, este estudio establece un marco de referencia claro para aplicar el entrenamiento incremental en diversos algoritmos, proporcionando una guía valiosa para futuras investigaciones y aplicaciones en la industria y la academia.

Investigación sobre algoritmos

- **Redes Neuronales Artificiales (ANN)**

Las Redes Neuronales Artificiales (ANN) de aprendizaje online representan una evolución significativa de las técnicas tradicionales, diseñadas para actualizar y mejorar sus modelos continuamente con nuevos datos. Esta capacidad es invaluable en entornos donde los datos son vastos o se generan de manera continua, como aplicaciones de streaming, finanzas o Internet de las Cosas (IoT).

Las ANN de aprendizaje online utilizan predominantemente el Stochastic Gradient Descent (SGD) para actualizar los pesos, procesando las muestras de entrada individualmente o en pequeños lotes (mini-batches), optimizando así la carga computacional y permitiendo una rápida adaptación a nuevos patrones. Para gestionar el overfitting se aplican técnicas como Dropout, Early Stopping y la regularización L1/L2. La elección de la función de pérdida varía según el tipo de problema, utilizando entropía cruzada para clasificación y error cuadrático medio para regresión.

Las tasas de aprendizaje adaptativas, como Adagrad, RMSprop o Adam, se ajustan según el rendimiento del modelo a lo largo del tiempo. Además, es fundamental realizar evaluaciones periódicas del modelo para monitorear su desempeño y detectar cualquier degradación o drift en los datos, utilizando técnicas de validación cruzada adaptativa o conjuntos de validación actualizados regularmente.

Las ANN de aprendizaje online son extremadamente útiles en escenarios donde los datos cambian con frecuencia, aplicándose en la detección de fraude en tiempo real, recomendaciones personalizadas en plataformas de streaming y análisis de sentimientos en redes sociales.

- **LGBM**

LightGBM (Light Gradient Boosting Machine) es un algoritmo de boosting de gradientes optimizado para eficiencia y efectividad, soportando el entrenamiento incremental, lo cual es esencial para aplicaciones con grandes volúmenes de datos. Utiliza técnicas de histograma para agrupar los valores continuos de las características en bins discretos, reduciendo significativamente el uso de memoria y el tiempo de cómputo. LightGBM es adecuado para el entrenamiento incremental debido a:

1. Entrenamiento Basado en Histograma: Facilita la actualización eficiente de los modelos con nuevos datos.
2. Manejo de Grandes Volúmenes de Datos: Capaz de manejar grandes datasets de manera eficiente.
3. API de Continuación de Entrenamiento: Permite continuar el entrenamiento desde un modelo guardado, evitando la necesidad de reiniciar el proceso desde cero.

El entrenamiento incremental en LightGBM se facilita cargando un modelo pre-entrenado y continuando el entrenamiento con nueva data, optimizando así los recursos computacionales y mejorando la capacidad de adaptación a cambios en los patrones de datos.

Hallazgos

A lo largo del proyecto, se lograron varios avances significativos que optimizaron tanto el proceso de entrenamiento de los modelos como la calidad de las predicciones. Primero, a través de la ingeniería de características, se generaron 50 columnas adicionales que enriquecieron el conjunto de datos. Este aumento en las características permitió una mejor representación de los datos y, por ende, un potencial aumento en la precisión de los modelos. Además, se utilizó la técnica SMOTE (Synthetic Minority Over-sampling Technique) para ajustar el balanceo de la variable objetivo a una proporción de 1:3. Este ajuste garantizó una representación más equitativa de las clases minoritarias, lo cual es crucial en conjuntos de datos desbalanceados, como es el caso de las transacciones fraudulentas.

El entrenamiento incremental se destacó como la solución más eficiente para ambos modelos, Redes Neuronales Artificiales (ANN) y LightGBM. En el caso de las ANN, se observó una diferencia notable en el tiempo de ejecución, siendo de casi 8 minutos para el entrenamiento secuencial y de casi 5 minutos para el incremental. A pesar de esta reducción en el tiempo, la efectividad del modelo no se vio comprometida. De hecho, los mejores resultados obtenidos para el entrenamiento incremental mostraron un F1 Score, precisión y exactitud del 100%, sin falsos positivos, lo que indica una alta confiabilidad en las predicciones. Para el modelo LightGBM, el entrenamiento incremental también demostró ser superior en términos de eficiencia temporal y recursos computacionales. Los resultados del entrenamiento incremental fueron consistentemente altos, con un F1 Score, precisión y exactitud del 100% en uno de los meses evaluados, y sin disminuir de un 96% en general. Esta consistencia resalta la ventaja del entrenamiento incremental, especialmente cuando se maneja un volumen creciente de datos, ya que evita la necesidad de reentrenar el modelo desde cero, reduciendo significativamente el tiempo de procesamiento.

En conclusión, los hallazgos de este proyecto subrayan la efectividad y eficiencia del entrenamiento incremental para modelos de aprendizaje automático en entornos con datos dinámicos y continuos. Tanto las ANN como LightGBM se beneficiaron de este enfoque, demostrando mejoras en el tiempo de entrenamiento y manteniendo una alta precisión en las predicciones. Estas ventajas hacen del entrenamiento incremental una estrategia esencial para aplicaciones en tiempo real y en escenarios con grandes volúmenes de datos.

Conclusión

Este proyecto ha evidenciado las ventajas del entrenamiento incremental en modelos de aprendizaje automático, especialmente en escenarios con datos dinámicos y en constante crecimiento.

1. Eficiencia en el Tiempo de Entrenamiento: El entrenamiento incremental redujo significativamente los tiempos de entrenamiento. En las ANN, el tiempo de entrenamiento se redujo de 8 a 5 minutos, mientras que en LightGBM, el enfoque incremental también mostró una mayor eficiencia temporal. Simbolizando también una alta eficiencia del uso de recursos computacionales.
2. Alta Precisión de los Modelos: Ambos modelos mantuvieron alta precisión utilizando el enfoque incremental. Las ANN lograron un F1 Score, precisión y exactitud del 100%, y LightGBM mantuvo una efectividad superior al 96%.
3. Balanceo de Datos con SMOTE: La aplicación de SMOTE ajustó el balanceo de la variable objetivo a una proporción de 1:3, mejorando la capacidad de los modelos para identificar transacciones fraudulentas.
4. Flexibilidad y Adaptabilidad: Los modelos demostraron ser flexibles y adaptables al incorporar nuevos datos, manteniéndose actualizados y precisos frente a cambios en los patrones de datos.

Referencias

Bishop, C. M. (n.d.). Pattern recognition and machine learning. microsoft.com. https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

Graham, M. (n.d.). “probabilistic machine learning”: A book series by Kevin Murphy. pml-book. https://probml.github.io/pml-book/

Hoi, S. C. H., Sahoo, D., Lu, J., & Zhao, P. (2018, October 22). Online learning: A comprehensive survey. arXiv.org. https://arxiv.org/abs/1802.02871

Ke, G., Meng, Q., Finely, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2019, August 6). LightGBM: A highly efficient gradient boosting decision tree. Microsoft Research. https://www.microsoft.com/en-us/research/publication/lightgbm-a-highly-efficient-gradient-boosting-decision-tree/

Laskov, P., Gehl, C., & Kruger, S. (2006). Incremental Support Vector Learning: Analysis, Implementation and Applications. Journal of Machine Learning Research . https://www.jmlr.org/papers/volume7/laskov06a/laskov06a.pdf

Utgoff, P. E. (1989, November). Incremental induction of Decision Trees - machine learning. SpringerLink. https://link.springer.com/article/10.1023/A:1022699900025

Washington, T. C. U. of, Chen, T., Washington, U. of, Washington, C. G. U. of, Guestrin, C., Ibm, Bosch, Amazon, Baidu, & Metrics, O. M. A. (2016, August 1). XGBoost: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and data mining. ACM Conferences. https://dl.acm.org/doi/10.1145/2939672.2939785