UNIVERSIDAD DEL VALLE DE GUATEMALA

Security Data Science

Sección 10

Proyecto#2

"Entrenamiento Incremental
en Modelos de Deep Learning y Machine Learning"

CHRISTOPHER EMANUEL ALEXANDER GARCIA PIXOLA 20541

GUATEMALA, Mayo 2024

Investigación Teórica

Entrenamiento Incremental en Aprendizaje Automático

El entrenamiento incremental, también conocido como aprendizaje online o continuo, es un paradigma de aprendizaje automático en el que los modelos se actualizan continuamente con nuevos datos a medida que están disponibles. Esto contrasta con el enfoque tradicional de entrenamiento por lotes, donde el modelo se entrena una vez con un conjunto de datos completo y luego se utiliza para hacer predicciones.

Beneficios del Entrenamiento Incremental:

- Eficiencia: El entrenamiento incremental puede ser más eficiente en términos de tiempo y recursos, ya que no es necesario esperar a que se recopilen todos los datos antes de comenzar el entrenamiento.
- Precisión: Puede ser más preciso, ya que el modelo puede adaptarse a los cambios en los datos a medida que ocurren.
- Escalabilidad: Puede ser más escalable, ya que el modelo se puede entrenar en un flujo de datos continuo.

¿Cómo funciona el Entrenamiento Incremental?:

Existen diferentes algoritmos para el entrenamiento incremental, pero la idea general es la misma: el modelo se actualiza con nuevos datos a medida que están disponibles. Esto se puede hacer de varias maneras, por ejemplo:

- Aprendizaje online: En este enfoque, el modelo se actualiza después de cada nuevo ejemplo de entrenamiento.
- Aprendizaje por lotes mini: En este enfoque, el modelo se actualiza después de un pequeño subconjunto de ejemplos de entrenamiento.
- Aprendizaje con ventanas deslizantes: En este enfoque, el modelo se entrena en un subconjunto de datos que se mueve a medida que se reciben nuevos datos.

Aplicaciones del Entrenamiento Incremental

El entrenamiento incremental tiene una amplia gama de aplicaciones, como: Detección de anomalías, reconocimiento de voz, visión por computador. Si bien el entrenamiento incremental tiene muchas ventajas, también tiene algunas limitaciones:

- Optimización: Puede ser más difícil optimizar los modelos entrenados incrementalmente, ya que los parámetros del modelo se actualizan continuamente.
- Tamaño del modelo: Los modelos entrenados incrementalmente pueden ser más grandes que los modelos entrenados por lotes, ya que almacenan el historial de actualizaciones de parámetros.

Metodología y Análisis de resultados

Los resultados muestran un rendimiento impresionante para ambos modelos (dentro del notebook se muestran las métricas de accuracy, matriz de confusión, precision, recall y f1-score), tanto el modelo de Red Neuronal Artificial (ANN) como el modelo LightGBM, con altos niveles de precisión y un rendimiento general sólido en la detección de transacciones fraudulentas.

Comenzando con el modelo ANN, se observa una precisión general del 99.49%, lo que indica una capacidad excepcional para predecir con precisión tanto las transacciones legítimas como las fraudulentas. La matriz de confusión revela que de las 3,770 transacciones fraudulentas, el modelo identifica correctamente el 91% de ellas, lo que refleja una alta sensibilidad o tasa de verdaderos positivos. Sin embargo, la precisión de la clase minoritaria (transacciones fraudulentas) es del 50%, lo que sugiere que una proporción significativa de las transacciones identificadas como fraudulentas podrían ser falsos positivos. (Se muestra en Imagen A)

Por otro lado, el modelo LightGBM también exhibe un rendimiento excepcional, con una precisión global del 99.44%. La matriz de confusión muestra que el modelo identifica el 88% de las transacciones fraudulentas, lo que indica una sensibilidad algo inferior en comparación con el modelo ANN. Sin embargo, la precisión de la clase minoritaria es del 55%, lo que sugiere una mejora en la capacidad de identificar transacciones fraudulentas con menor propensión a generar falsos positivos en comparación con el modelo ANN. (Se muestra en Imagen B)

En términos generales, ambos modelos logran un rendimiento impresionante en la detección de transacciones fraudulentas, con altos niveles de precisión y sensibilidad. Sin embargo, la elección entre los dos modelos puede depender de las necesidades específicas del negocio. Por ejemplo, si se prioriza la minimización de los falsos positivos, el modelo LightGBM podría ser preferible debido a su mayor precisión en la clase minoritaria. Por otro lado, si se valora más la sensibilidad o la capacidad para identificar la mayoría de las transacciones fraudulentas, el modelo ANN podría ser más adecuado debido a su mayor sensibilidad. En última instancia, una evaluación exhaustiva de los requisitos comerciales y las características del conjunto de datos podría ayudar a determinar el modelo más apropiado para implementar en un entorno de producción.

El éxito observado en la detección de transacciones fraudulentas puede atribuirse a varios factores, incluida la calidad y la preparación de los datos, así como las características intrínsecas de los modelos utilizados. En primer lugar, la eliminación de columnas innecesarias y la separación adecuada de características (X) y etiquetas (y) en el preprocesamiento de datos ayudaron a reducir la dimensionalidad y a mejorar la calidad de los datos de entrada para los modelos. Además, el equilibrio del conjunto de datos de entrenamiento mediante la técnica Synthetic Minority Over-sampling Technique (SMOTE) permitió mitigar el desequilibrio de clases, lo que posiblemente mejoró la capacidad de los modelos para capturar patrones en las transacciones fraudulentas.

En cuanto a la selección de modelos, la elección de una Red Neuronal Artificial (ANN) y LightGBM podría haber contribuido significativamente al éxito observado. Las ANN son conocidas por su capacidad para capturar relaciones no lineales y complejas en los datos, lo que las hace adecuadas para tareas de clasificación como la detección de fraudes. La capacidad de las ANN para aprender representaciones de características jerárquicas puede haber sido beneficiosa para identificar patrones sutiles en los datos que pueden indicar actividades fraudulentas.

Por otro lado, LightGBM es un algoritmo de aumento de gradiente que se destaca por su eficiencia computacional y su capacidad para manejar grandes conjuntos de datos con alta dimensionalidad. La capacidad de LightGBM para dividir el espacio de características en regiones más pequeñas mediante la construcción de árboles de decisión con un enfoque en los casos más difíciles puede haber sido crucial para identificar transacciones fraudulentas en un conjunto de datos con una alta proporción de transacciones legítimas.

```
Modelo ANN:
Accuracy: 0.9960065212873064

Matriz de Confusión:
[[734608   2580]
 [   379   3391]]

Reporte de Clasificación:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    737188
           1       0.57      0.90      0.70      3770

    accuracy                           1.00    740958
   macro avg       0.78      0.95      0.85    740958
weighted avg       1.00      1.00      1.00    740958
```

```
Modelo LightGBM:
Accuracy: 0.9944380037791899

Matriz de Confusión
[[734456   2732]
 [   451   3319]]

Reporte de Clasificación:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    737188
           1       0.55      0.88      0.68      3770

    accuracy                           1.00    740958
   macro avg       0.77      0.94      0.84    740958
weighted avg       1.00      1.00      1.00    740958
```

Imagen A y B: Métricas y matriz de confusión

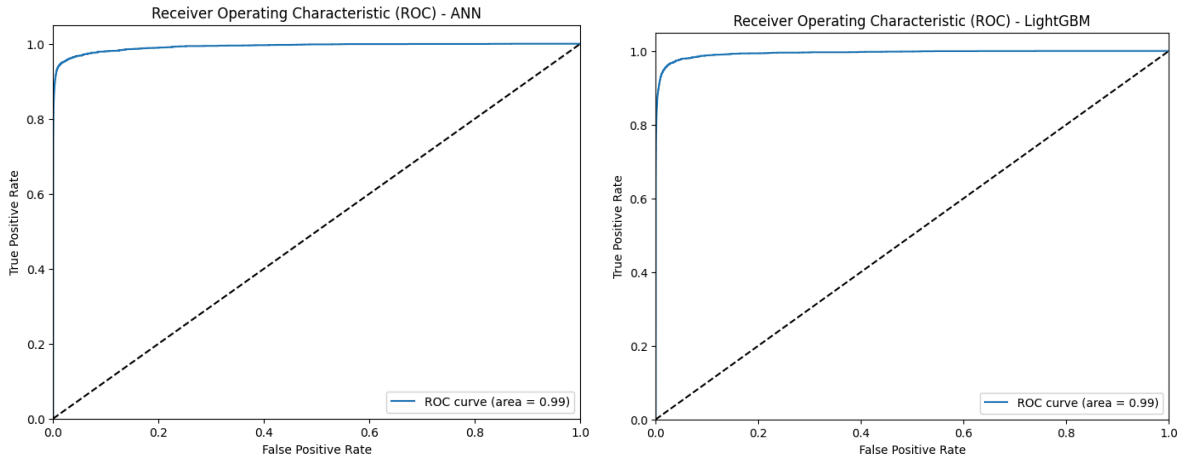

Imagen C y D: Curvas ROC

Conclusiones y recomendaciones

Conclusiones

- Rendimiento de los modelos: Los modelos de Red Neuronal Artificial (ANN) y LightGBM lograron un rendimiento muy bueno en la detección de transacciones fraudulentas, con altos niveles de precisión y sensibilidad. Esto demuestra la eficacia de las técnicas de preprocesamiento de datos y la selección de modelos adecuados para abordar el problema de desequilibrio de clases en conjuntos de datos financieros.

- Importancia del equilibrio de datos y la selección de características: El equilibrio del conjunto de datos de entrenamiento mediante SMOTE fue crucial para mejorar la capacidad de los modelos para detectar transacciones fraudulentas. Además, la eliminación de columnas innecesarias y la selección cuidadosa de características contribuyeron a la construcción de modelos más eficientes y precisos.

- Necesidad de una evaluación exhaustiva de modelos: Si bien tanto las ANN como LightGBM mostraron un rendimiento impresionante, es importante realizar una evaluación exhaustiva de diferentes modelos y técnicas para determinar la mejor solución en función de los requisitos comerciales y las características específicas del conjunto de datos.

Recomendaciones

- Optimización continua del modelo: A pesar del rendimiento observado, se recomienda continuar optimizando y ajustando los modelos para mejorar aún más su precisión y capacidad de generalización. Esto podría incluir la exploración de hiper parámetros, técnicas de selección de características y validación cruzada, todas estas a una mayor profundidad.

- Evaluación de hiper parámetros y división de datos: Aunque los modelos resultaron exitosos se considera prudente para un futuro evaluar la división de datos (desde la decisión de reducir o aumentar el dataset, hasta los períodos de tiempo utilizados) y también utilizar hiper parámetros para mejorar la capacidad del modelo para detectar fraudes.

- Interpretación y transparencia: Es importante comprender y comunicar de manera transparente cómo funcionan los modelos de detección de fraudes, especialmente en entornos regulados. Esto incluye la interpretación de características importantes, la explicación de decisiones del modelo y la documentación adecuada de procesos y resultados para garantizar la confianza y la aceptación por parte de todas las partes interesadas.

Bibliografía

- Incremental Learning for Artificial Neural Networks: https://www.nature.com/articles/s42256-022-00568-3
- Incremental Learning of Deep Neural Networks for Mobile Vision Detection: https://link.springer.com/article/10.1007/s10462-022-10294-2
- Incremental Learning of Convolutional Neural Networks with Bayesian Methods: https://arxiv.org/abs/1802.07329
- LightGBM Documentation: Incremental Learning: https://stats.stackexchange.com/questions/453540/how-does-lightgbm-deals-with-incremental-learning-and-concept-drift
- Incremental Learning with LightGBM for Click-Through Rate Prediction: https://dl.acm.org/doi/abs/10.1145/3569966.3570011
- XGBoost Documentation: Incremental Learning: https://shunya-vichar.medium.com/incremental-learning-in-xgboost-b3eac6135ce
- Incremental Learning for User Preference Prediction with XGBoost: https://ieeexplore.ieee.org/document/9315494
- Incremental Random Forests for Real-time Anomaly Detection: https://www.hindawi.com/journals/acisc/2022/1558381/
- Incremental Learning with Random Forests: https://stackoverflow.com/questions/44060432/incremental-training-of-random-forest-model-using-python-sklearn
- Incremental Support Vector Learning: http://www.stefan-rueping.de/publications/rueping-2001-b.pdf
- Incremental SVM Learning for Large-Scale Data Classification: https://medium.com/computers-papers-and-everything/incremental-learning-with-support-vector-machines-e838cd2d7691