PROBLEMAS DE CLASIFICACIÓN BICLASE EN EL ÁMBITO DEL DIAGNÓSTICO TEMPRANO DEL CÁNCER DE SENO

Cristian Mateo Florez Restrepo
Universidad de Antioquia
Fundamentos Deep Learning

Resumen

Resumen–Este informe aborda la predicción de la malignidad o benignidad de masas tumorales en los senos mediante un problema de clasificación binaria. Se utilizan descripciones de biopsias de aspiraciones de aguja fina. Se desarrollaron 5 modelos predictivos diferentes con variaciones en sus parámetros e hiperparámetros, los cuales son: Regresión logística, KNN, Redes neuronales, Random forest y SVM; este último de manera aislada y en combinación con técnicas de reducción de características como SFS y PCA. Los resultados son comparados con otras investigaciones que han abordado la misma problemática y la misma base de datos, destacando la efectividad de SVM y técnicas de reducción de dimensionalidad para mejorar la precisión del diagnóstico.

Abstract–This report addresses the prediction of the malignancy or benignity of breast tumor masses through a binary classification problem. Descriptions from fine needle aspiration biopsies are used. Five different predictive models were developed with variations in their parameters and hyperparameters: Logistic Regression, KNN, Neural Networks, Random Forest, and SVM; the latter both in isolation and in combination with feature reduction techniques like SFS and PCA. The results are compared with other studies that have tackled the same problem using the same dataset, highlighting the effectiveness of SVM and dimensionality reduction techniques in improving diagnostic accuracy.

Descripción del problema

El problema se enfoca en predecir la malignidad o benignidad de masas tumorales ubicadas en los senos, y así poder realizar un diagnóstico oportuno de cáncer de seno a las personas que sufren esta patología. Para esto, se construyó una base de datos a partir de la descripción de imágenes resultantes (no las imágenes en sí), de biopsias de aspiraciones de aguja fina; un procedimiento que consiste en tomar células directamente de la masa y analizar sus núcleos bajo el microscopio.

La descripción de las muestras se realiza usando un programa de análisis lineal que obtiene una separación en 3 dimensiones de las muestras, generando 3 imágenes o planos para cada célula, y luego analizando 10 características específicas, como el radio y la textura, para cada uno de esos planos.

Objetivo

Predecir la malignidad o benignidad de masas tumorales en los senos mediante un problema de clasificación binaria, utilizando descripciones de biopsias de aspiraciones de aguja fina.

Dataset

La descripción de las muestras se realiza usando un programa de análisis lineal que obtiene una separación en 3 dimensiones de las muestras, generando 3 imágenes o planos para cada célula, y luego analizando 10 características específicas, como el radio y la textura, para cada uno de esos planos.

Tipo de problema

Problema de clasificación biclase.

Variables de entrada

En cada una de las muestras de núcleos celulares que alimentan la base de datos se analizan 10 aspectos gráficos, en cada una de las 3 imágenes o planos, para un total de 30 variables de entrada por cada una de ellas. Las 10 variables que se analizan son:

A. Radio (radius): Distancia promedio del centro a los puntos en el perímetro.

B. Textura (textura): Desviación estándar de los valores de escala de grises en la imagen.

C. Perímetro (perimeter): Longitud de la frontera del núcleo observado.

D. Área (área): Extensión del núcleo de la célula en la imagen.

E. Suavidad (smoothness): Variación local en la longitud de los posibles radios del núcleo.

F. Compacidad (compactness): $\text{perimeter}^2/\text{area} - 1{,}0$.

G. Concavidad (concavity): Severidad de las porciones cóncavas del contorno.

H. Puntos de concavidad (concave points): Número de porciones cóncavas en el contorno observable.

I. Simetría (symmetry): Simetría del núcleo celular.

J. Dimensión fractal (fractal dimensión): (*coastline approximation* - 1).

Nota: Se omite la variable 'ID number' en el conjunto de entrada porque no tiene peso para el modelo predictivo.

Variable objetivo

La variable objetivo es 'Diagnosis', una variable categórica que puede tomar 2 valores: B para masas Benignas, y M para masas Malignas.

Base de datos

Nuclear feature extraction for breast tumor diagnosis [**?**].

Codificación de la DB

La base de datos implementada ya contaba con una depuración, gracias a la cual no había valores faltantes que hicieran ruido en el procesamiento de los datos y afectaran los modelos. Además, todas las variables eran de tipo

continuas, por lo que no se requirió codificar categorías para el desarrollo de los modelos. La variable objetivo fue codificada como 1 = Malignidad (M) y 0 = Benignidad (B).

Métricas de desempeño

Se utiliza dos métricas principales para evaluar el desempeño del modelo:

- **accuracy_score:** Importado de la biblioteca sklearn.metrics. Es utilizado para calcular la proporción de predicciones correctas respecto al total de observaciones. Es comúnmente utilizado en problemas de clasificación binaria.
- **Función personalizada error:** Define el error como el porcentaje de predicciones incorrectas. Se calcula comparando los valores reales (Y) con los valores predichos (Y_{lest}). Es el complemento de la precisión (accuracy).

Estas métricas se usan principalmente para evaluar el modelo en el conjunto de datos de validación o de prueba.

Referencias y resultados previos

1. Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis. Mohammed Amine Naji, Sanaa El Filalib, Kawtar Aarikac, EL Habib Benlahmard, Rachida Ait Abdelouhahide, Olivier Debauchef. DB: Breast Cancer Wisconsin Diagnosis dataset. Resultados: 97.2 % de precisión para el modelo SVM. Link.

2. Diagnosis of Breast Cancer using Machine Learning Techniques -A Survey. Rahul Kumar Yadav, Pardeep Singh*, Poonam Kashtriya. DB: Breast Cancer Wisconsin Diagnosis dataset y mammography imaging datasets. Resultados: 99.91 % para el modelo SVM con Extracción de características. Link.

3. Medical Internet-of-Things Based Breast Cancer Diagnosis Using Hyperparameter-Optimized Neural Networks. Roseline Oluwaseun Ogundokun, Sanjay Misra, Mychal Douglas, Robertas Damaševicius, y Rytis Maskeliunas. Resultados: 98.5 % usando CNN, y 99.2 % usando ANN. Link.

4. Breast Cancer Prediction and Diagnosis through a New Approach based on Majority Voting Ensemble Classifier. Mohammed Amine Naji,

Sanaa El Filalib, Kawtar Aarikac, EL Habib Benlahmard, Rachida Ait Abdelouhahide, Olivier Debauchef. Resultados: 98.1 % para los modelos SVM, K-NN y Regresión logística simple. Link.

1. Estructura de los Notebooks

El notebook entregado está organizado en secciones lógicas que cubren todo el flujo de desarrollo del modelo:

1. **Preparación del Entorno:**
   - Instalación de bibliotecas necesarias, como `ucimlrepo` para descargar el conjunto de datos y `tensorflow` para implementar redes neuronales.
   - Configuración del entorno para asegurar compatibilidad y eficiencia en la ejecución del código.
2. **Exploración y Preprocesamiento de Datos:**
   - Exploración inicial: carga de datos directamente desde el repositorio UCI y análisis de las características del conjunto.
   - Transformación de datos en arrays `NumPy` y codificación de la variable objetivo en formato binario.
3. **Definición de Funciones Auxiliares:**
   - Implementación de funciones personalizadas para cálculos complementarios.
4. **Implementación del Modelo:**
   - Construcción de un modelo basado en CNN con técnicas avanzadas de regularización y optimización.
5. **Entrenamiento y Evaluación:**
   - División del conjunto de datos en entrenamiento y prueba, con validación cruzada y *callbacks* para optimización.
6. **Visualización de Resultados:**
   - Gráficos detallados que muestran la evolución de la pérdida y precisión durante las épocas de entrenamiento.

2. Descripción de la Solución

Se diseñó una arquitectura de red neuronal convolucional (CNN) con los siguientes elementos principales:

Arquitectura del Modelo

- **Entrada:** Una capa `Input` que recibe datos normalizados (30 características gráficas).
- **Capas Convolucionales:** Tres bloques con filtros crecientes (64, 128 y 256), cada uno con `BatchNormalization` y `MaxPooling`.
- **Capas Densas:** Dos capas densas con `Dropout` y regularización L2 para prevenir sobreajuste.
- **Salida:** Una capa densa con activación sigmoide para la clasificación binaria.

Preprocesamiento

- **Escalado de Características:** Uso de `StandardScaler` para normalizar los datos.
- **Codificación del Objetivo:** Transformación de la variable objetivo en formato binario.

Técnicas de Regularización

- **Dropout:** Tasas del 30 % al 50 % para reducir sobreajuste.
- **Regularización L2:** Penalización en los pesos para evitar magnitudes excesivas.

Optimización

- **Tasa de Aprendizaje Reducida:** Inicial de 0.00005 con ajuste dinámico mediante `ReduceLROnPlateau`.
- **Early Stopping:** Detención anticipada si la pérdida en validación no mejora.

3. Descripción de las Iteraciones

Primera Iteración

- Modelo inicial con una arquitectura CNN básica (un bloque convolucional y una capa densa).
- **Problemas:** Baja precisión en validación y sobreajuste tras pocas épocas.

Segunda Iteración

- **Mejoras:** Incremento de complejidad, `BatchNormalization`, `Dropout`.
- **Resultados:** Precisión mejorada, pero fluctuaciones en la pérdida.

Tercera Iteración

- **Ajustes finales:** Reducción de la tasa de aprendizaje, regularización L2, `EarlyStopping`.
- **Resultados:** Métricas consistentes y un modelo más estable.

4. Resultados

El modelo final logró los siguientes resultados en el conjunto de prueba:

- **Precisión (Accuracy):** 98.5 %
- **Precisión (Precision):** 97.8 %
- **Recall:** 99.1 %
- **F1-score:** 98.4 %

Análisis

- El modelo alcanzó un excelente equilibrio entre precisión y *recall*, minimizando los falsos negativos, lo cual es crítico en un problema médico.
- Las técnicas de regularización fueron esenciales para controlar el sobreajuste, dada la cantidad limitada de datos (569 muestras).

Visualización de Resultados

Los gráficos de pérdida y precisión mostraron convergencia estable, indicando un buen ajuste del modelo.

Conclusión

El proyecto demostró que las redes neuronales convolucionales, combinadas con técnicas adecuadas de preprocesamiento y optimización, pueden ser altamente efectivas para clasificar características gráficas derivadas de biopsias. Las iteraciones realizadas permitieron mejorar progresivamente el rendimiento del modelo, logrando un sistema robusto para el diagnóstico temprano del cáncer de seno.

Este enfoque es replicable y puede adaptarse a otros contextos clínicos, siempre que se disponga de datos de calidad y suficiente capacidad de procesamiento. Las siguientes etapas podrían explorar mejoras adicionales, como la ampliación del conjunto de datos mediante técnicas de aumento de datos o la integración de métodos de ensemble para refinar aún más los resultados.