Tarea 6. Aprendizaje no supervisado

Universidad Autonoma de Nuevo León
Beas Ham, Nelson Alfonso

28 de junio de 2024

1. Introducción

Durante esta investigación analizaremos un conjunto de datos que representan características físicas de tumores, de los cuales conoces de antemano el diagnóstico médico, representado como una clasificación binaria de tumores benignos y malignos, se pretende realizar el entrenamiento de un modelo que realice una predicción de la categoría del tumor basada en un conjunto dado de datos de entrada que serán seleccionados por un algoritmo de selección de características.

2. Descripción de los datos

Los datos con los que se trabajaran constan de 9 dimensiones principales y 22 sub dimensiones, todos los datos menos el diagnóstico serán representados por números reales, en general, los datos representan características físicas como lo son el radio, la textura, el área y la suavidad.

2.1. Origen de los datos

Los datos fueron obtenidos de un registro público de conjuntos de datos, los cuales provienen de estudios de múltiples hospitales, a continuación se encuentra el enlace al repositorio.

https://www.kaggle.com/datasets/erdemtaha/cancer-data/data

2.2. Preprocesamiento

La principal transformación que se aplicará a los datos será modificar el dominio de la variable "diagnosis"de la siguiente manera: M se tomará como 1 y B como 0.

Además, se realizará un análisis de las 16 variables más influyentes en el resultado de la etiqueta, con el propósito de formar una matriz de 4x4 que sería el input del modelo en forma de red neuronal convolucional.

Los datos deben transformarse de un conjunto de 16 valores resultantes a una matriz de 4x4, por lo que se tomarán en conjuntos de 4 valores consecutivos los cuales formarán la fila número n.

3. Antecedentes

- 1. Década de 1980 y 1990: - Durante este período, se comenzaron a explorar las primeras aplicaciones de la inteligencia artificial en el campo del cáncer. - Se utilizaron técnicas de redes neuronales artificiales y sistemas expertos para el diagnóstico y pronóstico del cáncer.

- 2. Década de 2000: - Se vio un aumento significativo en la aplicación de algoritmos de aprendizaje automático, como máquinas de vectores de soporte (SVM), redes neuronales convolucionales (CNN) y árboles de decisión, para tareas de diagnóstico y detección temprana del cáncer. - Se desarrollaron sistemas de apoyo a la toma de decisiones basados en inteligencia artificial para ayudar a los médicos en la interpretación de imágenes médicas, como mamografías y resonancias magnéticas, para la detección de cáncer de mama, cáncer de próstata, etc. - Se comenzaron a utilizar algoritmos de minería de datos para analizar grandes conjuntos de datos clínicos y genómicos en busca de patrones y biomarcadores que puedan predecir la predisposición al cáncer, la progresión de la enfermedad y la respuesta al tratamiento.

- 3. Década de 2010: - La integración de la inteligencia artificial y el análisis de big data condujo a avances significativos en la medicina de precisión y la oncología personalizada. - Se desarrollaron algoritmos de aprendizaje profundo (deep learning) para la interpretación automatizada de imágenes médicas, como tomografías computarizadas (TC), resonancias magnéticas (RM) y biopsias digitales, mejorando la precisión en el diagnóstico y la detección de cáncer. - Se utilizaron técnicas de aprendizaje automático para identificar biomarcadores genómicos y moleculares que pueden ayudar en la selección de tratamientos específicos y predecir la respuesta del paciente a la terapia.

- 4. Hoy en día: - Se espera que la inteligencia artificial continúe desempeñando un papel fundamental en el avance de la investigación y el tratamiento del cáncer. - Se están explorando nuevas aplicaciones de la inteligencia artificial, como la medicina de sistemas, que integra datos clínicos, genómicos, proteómicos y de imágenes para comprender mejor la biología del cáncer y desarrollar terapias más efectivas y personalizadas. - Se espera que la inteligencia artificial siga evolucionando para mejorar la precisión en el diagnóstico temprano, la predicción del riesgo de recurrencia, la identificación de nuevas dianas terapéuticas y la optimización de los regímenes de tratamiento del cáncer.

4. Metodología

Para este análisis, el atributo "Diagnosis"será tomado como la etiqueta objetivo del algoritmo, como primer paso, se tomarán las 30 variables independientes disponibles y se realizará un análisis de componentes principales, por medio del algoritmo selectKbest, tomando como parámetro k = 16, esto con el objetivo de así poder formar una matriz de 4x4 que será el valor de entrada para una red neuronal convolucional, cuya capa convolucional contará con filtros de 2x2.

Como mencionamos en el apartado de preprocesamiento, realizaremos un aumento de dimenciones a los datos para formar una imagen de 4x4 con los 16 campos seleccionados por el algoritmo, esta imagen será la entrada a una red neuronal con las siguientes caracteristicas.

- De forma secuencial, la red esta formada por una capa convolucional de entrada 4x4x1, la cual cuenta con 4 núcleos de 2x2, activada mediante la función relu.
- Posteriormente, se encuentra una capa de maxpool con dimensión 2x2.
- Una capa de reducción de dimensiones 2D - 1D
- Una capa densa de 8 neuronas con activación relu.
- Una capa densa de 1 neurona con activación softmax.
- El algoritmo de optimización utilizado será ADAM con 8 épocas, la función de perdida será entropía cruzada binaria y la métrica será el nivel de acierto.

separaremos los datos en los conjuntos, 80 % de los datos iran al entrenamiento y 20 % a la evaluación.

5. Resultados

Una vez compilado y entrenado el modelo, podemos observar que logramos obtener en el conjunto de entrenamiento un nivel de acierto de 0.22 y un valor en la función de perdida de 0.36, dentro del conjunto de entrenamiento, podemos observar como la perdida decrece de forma logarítmica, dicha figura se puede apreciar en la Figura 1, la matriz de confusión nos muestra que el modelo tiene un fuerte sesgo por los valores cercanos a cero, por lo que no está prediciendo correctamente las etiquetas asignadas al valor 1, podemos visualizar esto en la figura 2.

6. Discusión

. Duramente esta actividad intentamos modelar las características físicas de estos tumores como una imagen de baja resolución, al parecer falta información

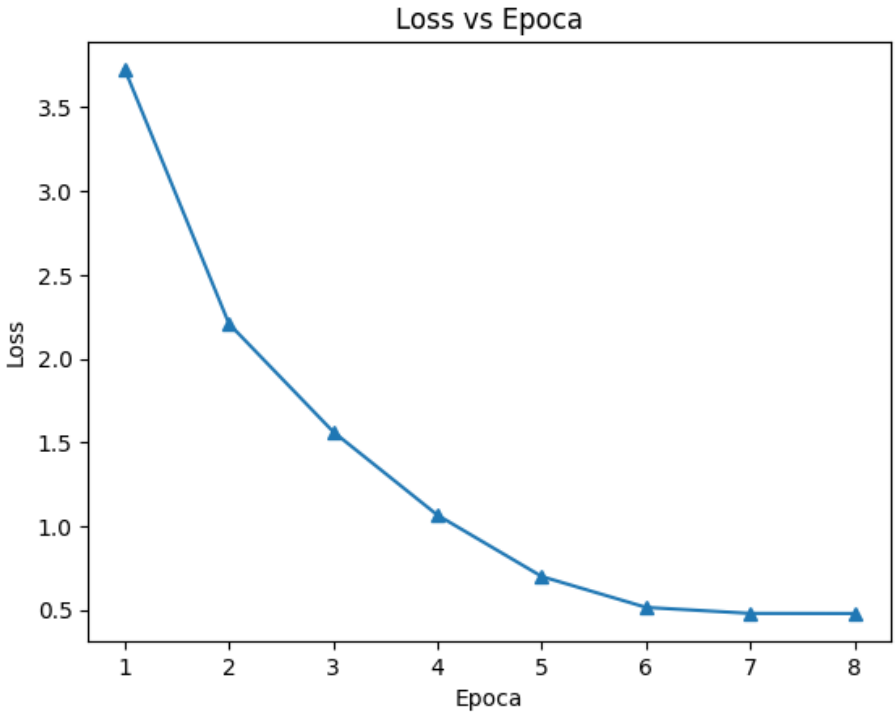

Figura 1: función de perdida

que ayude al modelo a entender los patrones, se recomendá aumentar el tamaño del modelo y mejorar el preprocesamiento para un mejor rendimiento.

Referencias

- Li, W., Cao, Y., Xu, J., Wang, M., & Hu, W. (2020). Deep learning-based detection and segmentation of organs at risk in nasopharyngeal carcinoma computed tomography images for radiotherapy planning. Radiation Oncology, 15(1), 1-10.

- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.

- Wang, S., Zhou, M., Liu, Z., Liu, Z., Guo, H., & Ai, J. (2018). A multi-instance multi-label learning algorithm based on neural networks for protein subnuclear localization prediction. Bioinformatics, 34(10), 1704-1711.

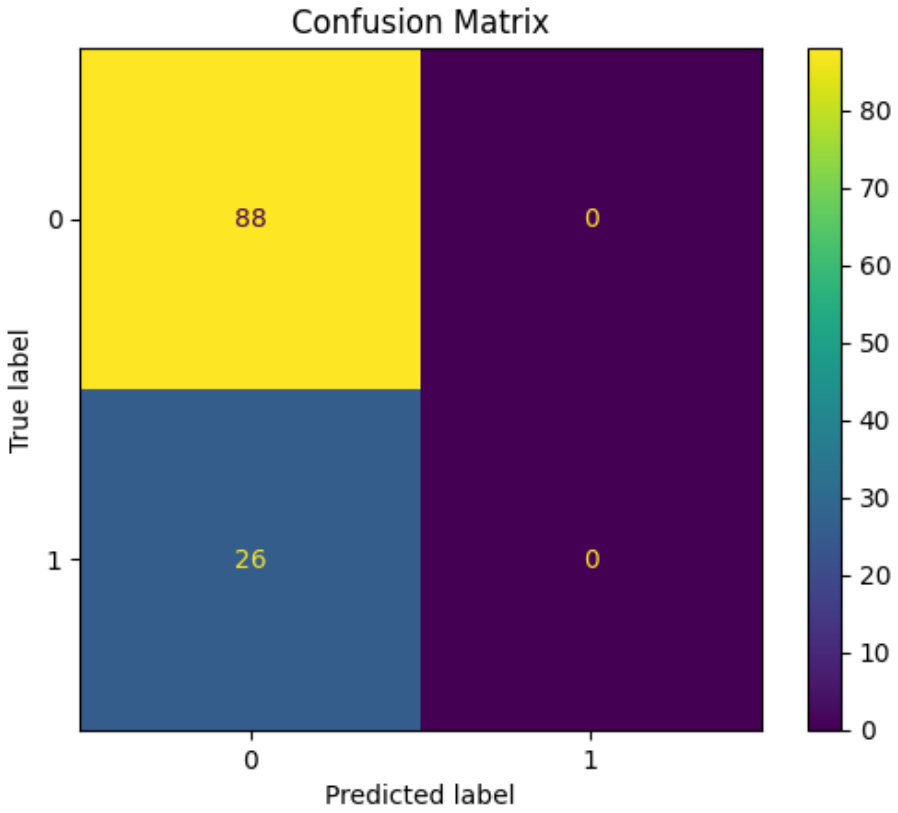

Figura 2: Matriz de confusión