Proyecto aprendizaje automático

Universidad Autonoma de Nuevo León
Beas Ham, Nelson Alfonso

22 de julio 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, se tendrá en competición un conjunto de tres algoritmos los cuales simultáneamente generarán un resultado, asignado así al dato la etiqueta más votada.

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 algoritmo supervisado

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.

2.3. Preprocesamiento algoritmo no supervisado

Se aplicará a los datos una moficicación dell dominio de la variable "diagnosis"de la siguiente manera: M se tomará como 1 y B como 0, ademas de su seperación del conjuto de datos de entranamiento.

Además, se realizará un análisis de las 8 variables seleccionadas por un algorimto de PCA, las cuales serán las cantidatas para la generación de los closters.

2.4. Preprocesamiento del arbol

Se aplicará a los datos una moficicación dell dominio de la variable "diagnosis"de la siguiente manera: M se tomará como 1 y B como 0, ademas de su seperación del conjuto de datos de entranamiento.

Además, se realizará un análisis de las 16 variables seleccionadas por un algorimto de PCA, las cuales serán las cantidatas para la generación de los nodos del arbol.

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. Marco teórico

En esta sección agruparemos los componentes teoricos que conformaran las tecnicas de analisis de desempeño para nuestros algoritmos seleccionados, a continuacion listaremos las tecnicas seleccionadas para este proyecto.

4.1. Metrica de desempeño del algoritmo no supervisado

Inercia (Within-Cluster Sum of Squares, WCSS)

La **inercia** es una métrica que mide la compacidad de los clusters formados por el algoritmo k-means. Se define como la suma de las distancias al cuadrado de cada punto a su centroide más cercano.

Fórmula

$$\text{Inercia} = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

donde:

- k es el número de clusters.
- C_i es el conjunto de puntos en el cluster i.
- μ_i es el centroide del cluster i.

- $\|x - \mu_i\|^2$ es la distancia euclidiana al cuadrado entre el punto x y el centroide μ_i.

Interpretación

- Una menor inercia indica clusters más compactos y bien definidos.
- Se busca minimizar la inercia al ajustar el número de clusters k.

4.2. Metrica de desempeño del algoritmo supervisado

La exactitud se define como la proporción de instancias correctamente clasificadas sobre el total de instancias evaluadas.

Definición

$$\text{Exactitud} = \frac{\text{Número de Predicciones Correctas}}{\text{Número Total de Predicciones}}$$

Interpretación

- Una exactitud alta indica que el modelo está haciendo muchas predicciones correctas.
- Una exactitud baja indica que el modelo está haciendo muchas predicciones incorrectas.

La exactitud es fundamental en la evaluación de redes neuronales convolucionales, proporcionando una medida directa de su capacidad para clasificar correctamente datos de entrada.

4.3. Metrica de desempeño del algoritmo de arbol de clasificación

El error de clasificación es una métrica que mide la proporción de instancias mal clasificadas respecto al total de instancias evaluadas.

Definición

El error de clasificación se calcula como:

$$\text{Error de Clasificación} = \frac{\text{Número de predicciones incorrectas}}{\text{Número Total de Predicciones}}$$

donde:

- **Número de predicciones incorrectas**: Es la cantidad de instancias que fueron clasificadas incorrectamente por el modelo.
- **Número Total de Predicciones**: Es el total de instancias evaluadas por el modelo.

Interpretación

Una menor tasa de error de clasificación indica que el modelo está haciendo menos predicciones incorrectas, lo cual es deseable para evaluar el rendimiento de algoritmos de clasificación.

5. Metodología

Para este proyecto, tomaremos un conjunto de tres algoritmos en competición para definir la etiqueta que se asignará al dato de entrada, haciendo uso de un algoritmo supervisado, en forma de red neuronal convolucional, un algoritmo no supervisado de agrupación k-means y un árbol de clasificación binaria.

5.1. Metodología algoritmo supervisado

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.

5.2. Metodología algoritmo no supervisado

Para este análisis, el atributo "Diagnosis"no le será proporcionado al algoritmo, haciendo uso de la librería pandas y sklearn del lenguaje de programación python a continuación describiremos las bases del algoritmo candidato para la clasificación.

El algoritmo K-Means es un método de agrupamiento (clustering) que divide un conjunto de datos en K grupos (clusters) basados en la similitud de las características de los datos. Funciona de la siguiente manera:

1. **Inicialización:** Selecciona aleatoriamente K puntos del conjunto de datos como los centroides iniciales de los clusters.
2. **Asignación de puntos:** Para cada punto en el conjunto de datos, calcula la distancia entre el punto y todos los centroides. Asigna el punto al cluster cuyo centroide está más cercano.
3. **Actualización de centroides:** Recalcula los centroides de los clusters como el promedio de todos los puntos asignados a ese cluster.
4. **Repetición:** Repite los pasos 2 y 3 hasta que no haya cambios significativos en la asignación de puntos a los clusters o se alcance un número máximo de iteraciones.

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

5.3. Metodología algoritmo de arbol binario

Para este análisis, el atributo "Diagnosis"será tomado como la etiqueta objetivo del algoritmo, despues, como se mencionó en el apartado de pre-procesamiento, se tomanrán las 16 variables más importantes según el PCA, con las cuales se contruiran los nodos de la siguiente manera, a continuación describiremos la metodologia general para contruir un arbol de estas caracteristicas.

1. **Seleccionar la Mejor División**:
   - Evaluar todas las variables y umbrales posibles.
   - Calcular la pureza (índice de Gini, entropía, etc.) para cada división.
   - Seleccionar la variable y el umbral con la mejor pureza.
2. **Dividir el Conjunto de Datos**:
   - Usar la mejor variable y umbral para dividir los datos en dos subconjuntos.
   - Crear nodos hijos para cada subconjunto.
3. **Asignar Etiquetas a las Hojas**:

- Si un nodo contiene datos de una sola clase o se ha alcanzado la profundidad máxima, asignar la etiqueta de la clase mayoritaria.
- Si no, aplicar recursivamente los pasos 1 y 2 a los nodos hijos.

4. **Repetir Recursivamente Hasta Completar el Árbol**:
   - Continuar dividiendo cada nodo hijo recursivamente.
   - Detener cuando todos los nodos sean hojas o se cumplan los criterios de detención.

6. Resultados

Una vez entrenados los modelos, procedemos a generar un nuevo dato de prueba, el cual contendrá todas las 8 o 16 variables requeridas por cada modelo con un valor igual a 1, exceptuando por el radio que tendrá un valor de 10, los modelos han presentado la misma respuesta para su etiqueta, la cual predice que los valores provienen de un tumor que es benigno.

7. Discusión

. Dentro de los tres modelos seleccionados, el que obtuvo menor rendimiento fue el supervisado, se barajan alternativas para su mejora como el uso de optimizadores pre-entrenados, o inclusive técnicas para generar una imagen de más resolución, ya que una posible causa de su mal rendimiento sería que una imagen de 4x4 no permite mostrar los detalles correctamente del fenómeno, también existe la posibilidad de mejora acomodando de distinta forma la posición de cada dato en la imagen.

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.