Simple Random Forest Algorithm

Tabla de contenido

Raimundo Flores Cabra
dpto. Ciencias de la Computación e Inteligencia Artificial
Universidad de Sevilla
Sevilla, España
Correo electrónico UVUS:
raiflocab@alum.us.es
Correo electrónico de contacto:
raimundokarate98@gmail.com

Vicente López Vázquez
dpto. Ciencias de la Computación e Inteligencia Artificial
Universidad de Sevilla
Sevilla, España
Correo electrónico UVUS:
viclopvaz1@alum.us.es
Correo electrónico de contacto:
vlopezvazquez3@gmail.com

El objetivo de nuestro trabajo ha sido construir una versión simplificada del algoritmo de Bosques Aleatorios. Pero ¿en qué consiste este algoritmo? Un bosque aleatorio es un conjunto de árboles de decisión entrenados a partir de un conjunto de datos de entrenamiento con un conjunto aleatorio de atributos seleccionados para cada uno y tras la aplicación de Bootstrapping como técnica de muestreo. Esto nos permite construir un conjunto de árboles de decisión partiendo de un solo conjunto de datos de entrenamiento, ya que dichos datos son difíciles de conseguir y no muy abundantes. Por esta razón, se aplican técnicas de muestreo, en este caso Bootstrapping con reemplazo, para dar lugar a un nuevo conjunto de datos para cada árbol de decisión.

Esta técnica facilita el desarrollo de sistemas de predicción fiables y con mayor capacidad de predicción que un solo árbol de decisión. Debemos destacar que, a lo largo de este proyecto hemos comprendido que los datos con los que trabajamos no siempre son suficientes y tenemos que adaptarnos a ellos; que las técnicas de muestreo nos facilitan el trabajo con conjuntos de datos pequeños y con la construcción de varios árboles de decisión; y que un solo árbol no es tan eficaz como un conjunto promediado de ellos.

1. Introducción

La Inteligencia Artificial [1] tiene numerosos usos actualmente, desde sistemas de planificación automática o sistemas de decisión, a reconocimiento de la escritura o del habla. Nuestro caso trata sobre los árboles de decisión [2], un ejemplo de Aprendizaje Supervisado [3].

Un árbol de decisión es un modelo de predicción construido a partir de un conjunto de datos, de donde se extraen una serie de reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de manera sucesiva, para la resolución del problema. En la siguiente figura podemos ver la estructura de un árbol de decisión:

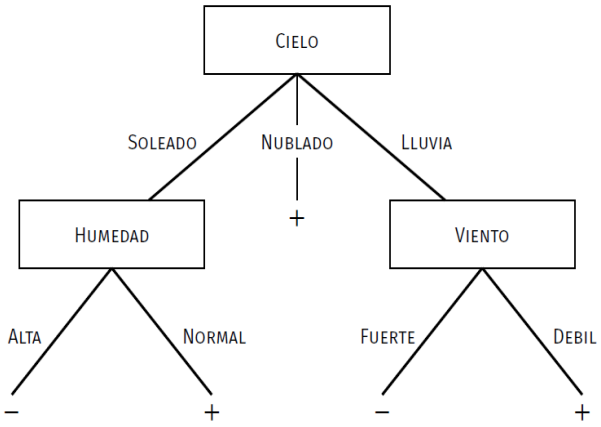

Figura 1 Árbol de decisión con 3 nodos de decisión.

Podemos ver que el árbol clasificará los distintos ejemplos teniendo en cuenta como primera forma de clasificación el atributo "Cielo". Dependiendo del valor de dicho atributo, el siguiente atributo usado para clasificar el ejemplo será "Humedad" o "Viento" o en caso de que el valor de "Cielo" sea nublado la clasificación es directa.

A grandes rasgos, nuestro proyecto trabaja con dichos árboles de decisión. Pero se complica al introducir técnicas de muestreo como Bootstrapping [4] y Random Subspace Method [5]. Además, trabajaremos con varios árboles con el objetivo de aumentar la precisión calculando la salida promedia. La aplicación de técnicas de muestreo para obtener nuevos conjuntos de datos a partir de uno y el uso de varios árboles entrenados con diferentes conjuntos de datos da lugar a una técnica llamada Random Forest (Bosque Aleatorio) [6].

El uso de Random Forest aumenta la precisión que el uso de un solo árbol de decisión proporciona y facilita el trabajo con conjuntos de datos pequeños o difíciles de conseguir. Por eso su uso está muy extendido y aceptado.

Como conclusión, nuestro proyecto ha consistido en diseñar y construir una forma simple del algoritmo del Random Forest, compararlo con el uso de árboles individuales, profundizar en

el uso de las técnicas de muestreo arriba citadas y experimentar con distintos parámetros para llegar a los mejores valores de predicción posibles.

2. Preliminares

Antes de entrar de lleno en el trabajo, pasaremos a numerar y explicar en que consisten las técnicas que hemos usado durante el desarrollo del proyecto. Primero hablaremos de la técnica básica en la que se basa nuestro trabajo, para después entrar en las técnicas de muestreo utilizadas y finalmente en la técnica que se crea como conjunto de las anteriores:

- **Técnicas básicas:**
    - **Aprendizaje supervisado [3]:** El aprendizaje supervisado es una técnica usada para clasificar datos o ejemplos a partir de un conjunto de datos de entrenamiento. Los conjuntos de entrenamiento son vectores de datos donde cada dato está formado por unos atributos y por una clasificación o valor numérico.

      El objetivo del aprendizaje supervisado es el de dar lugar a una función de decisión que pueda responder correctamente con un margen de error a unos datos de entrada según el entrenamiento dado. Los árboles de decisión son un ejemplo de aprendizaje supervisado.
    - **Árboles de decisión [2]:** Un árbol de decisión es una técnica de aprendizaje automático, y como hemos podido ver arriba, forma parte del aprendizaje supervisado ya que recibe un conjunto de datos de entrenamiento con el fin de entrenar al árbol para que pueda responder con una probabilidad alta de acierto frente a nuevos datos de entrada. Esta técnica es en la que se basa nuestro trabajo, pero la complicamos al aplicar distintas técnicas de muestreo al conjunto de datos de entrenamiento.

- **Técnicas de muestreo:**

- **Bootstrapping [4]:** Es una técnica de muestreo usada cuando tenemos un conjunto de datos limitado y no podemos obtener más. Dado un conjunto de datos de N filas, crearemos un nuevo conjunto de datos del mismo tamaño a partir de este.

  Dicho nuevo conjunto se crea recorriendo el conjunto de datos base; haciendo una selección con probabilidad de determinados ejemplos durante el recorrido; dichos ejemplos seleccionados sustituirlos por una copia de otro ejemplo cualquiera seleccionado con probabilidad. Así hasta generar un nuevo conjunto de datos de tamaño N o de un tamaño dado por el usuario.

  Es una técnica muy útil si, como hemos dicho antes, no contamos con un conjunto de datos abundante y tenemos necesidad de más datos que los que podemos conseguir.
- **Random Subspace Method [5]:** Otra técnica de muestreo usada en el ensamble de modelos. Se basa en la selección aleatoria de determinadas características del conjunto de datos de entrenamiento. El objetivo de esta técnica es que cada modelo de predicción se entrene con un conjunto de características concretas en vez de con el conjunto entero de estas, es decir, que cada modelo entrenado se especialice en un conjunto más pequeño de características para aumentar su precisión finalmente ensamblando todos los modelos y promediando la respuesta.

- **Técnica de ensamble:**
  - **Random Forest (Simplified) [6]:** Es la técnica de ensamble que hemos usado. Hemos desarrollado una versión simplificada de Random Forest o Bosque Aleatorio traducido.

    Consiste en: partiendo de un conjunto de datos de entrenamiento y un número de árboles a entrenar, para cada árbol se aplican las técnicas de muestreo arriba descritas al conjunto de datos y se entrena el árbol. Tras entrenarlos todo, esta técnica devuelve un conjunto de árboles (Bosque) entrenados a partir de un mismo conjunto de datos, pero cada uno especializado en un conjunto de características (Aleatorio). Para las predicciones se promediarían

las salidas del conjunto de árboles. Esta técnica proporciona una precisión mayor que un único árbol de decisión.

3. Metodología

Referencias

[1] https://es.wikipedia.org/wiki/Inteligencia_artificial
[2] https://es.wikipedia.org/wiki/%C3%81rbol_de_decisi%C3%B3n
[3] https://es.wikipedia.org/wiki/Aprendizaje_supervisado
[4] https://es.wikipedia.org/wiki/Bootstrapping_(estad%C3%ADstica)
[5] https://en.wikipedia.org/wiki/Random_subspace_method
[6] https://es.wikipedia.org/wiki/Random_forest
[7]

Figuras