Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación

Máquinas de aprendizaje

Reporte: Clasificación con árboles de decisión

Docente: Abraham Sánchez López

Alumno	Matrícula
Taisen Romero Bañuelos	202055209

Clasificación con árboles de decisión

Creo que la clasificación de árboles tiene su encanto no sólo en su efectividad, si no también en su semejanza al proceso natural de pensamiento de las personas para la toma de decisiones. Es curioso como métodos de este estilo destacan incluso en la biología de nuestro cerebro, quizá explorar métodos semejantes a nuestros procesos de pensamiento sea una buena estrategia ya que de cierta forma han pasado la prueba de la evolución (o adaptabilidad, mejor dicho).

En fin, para aplicar estos métodos tenemos a C5.0 y otros algoritmos que son casi hermanos primos de éste, por lo que las cosas que se mencionan aplican también para esos otros algoritmos. Primero seleccionan las características más relevantes de los datos para dividirlas consecutivamente basándose en las características que se pueden extraer de esas características, haciendo grupos de datos cada vez más homogéneos (un poco parecido a clustering si se ve con detenimiento).

Pese a ser efectivo peca de tener talones de aquiles, aunque varios se pueden resolver con algunas técnicas. La debilidad más relevante es que pueden sesgarse hacia divisiones con una gran cantidad de niveles. También otra es que se puede sub-ajustar o sobre ajustar con facilidad, y de hecho una técnica aplicada es primero sobre ajustar el árbol para después podarlo, pero el problema que no se menciona sobre esto es que puede resultar computacionalmente costoso, lo cual es un dato a tener en cuenta. Aunque eso no quita lo bonito y fácil que es interpretar los resultados de un árbol construido, a diferencia de otros modelos.

Para poder tratar el problema de saber cómo dividir los datos se aplica un concepto de la física llamado “entropía”. La entropía se puede definir como la aleatoriedad sin correlación entre datos, entonces, para poder trabajar lo que se hace es buscar volver esos datos en datos más puros (homogéneos). A medida que una clase domina a otra podemos reducir la entropía, entonces, con esta “dominancia” podemos calcular qué tan dominante tiene que ser una clase para reducir la entropía.

Otro detalle a mencionar es que todos estos cálculos que se mencionan consideran características nominales, así que hay que verificar el tipo de datos antes de hacer un árbol, ya que evidentemente puede provocar errores. Pero esto me recuerda a que otro problema de los árboles es que si cambiamos un poco los datos de entrenamiento podemos alterar mucho la lógica de decisión, casi como un efecto mariposa.