Práctica 5: Reglas clasificación: creación y evaluación de hipótesis con distintos algoritmos

Asignatura: Técnicas de aprendizaje automático (TAA)
Profesor: Teodoro Calonge
Alumno: Miguel Chaveinte García

1. Descripción conjunto de datos

- Contact-lenses: Conjunto de datos utilizados para determinar el tipo de lentes que necesita una persona en función de su edad y los diferentes problemas de visión con los que cuente.
  - → https://archive.ics.uci.edu/ml/datasets/Lenses
  - → 24 instancias
  - → 4 atributos (1+ clase)
  - → 3 clases
- Iris: Conjunto de datos multivariante que contiene datos que cuantifican la variación morfológica de la flor Iris de tres especies (setosa,virginica y versicolor).
  - → https://archive.ics.uci.edu/ml/datasets/iris
  - → 150 instancias (50 de cada clase)
  - → 4 atributos (numéricos)
  - → 3 clases
- Soybean: Definido en la Entrega 4 (https://aulas.inf.uva.es/mod/assign/view.php?id=18768)
  - → https://archive.ics.uci.edu/ml/datasets/Soybean+(Large)
  - → 683 instancias
  - → 35 atributos (todos nominales)
  - → 19 clases
- Vote: Definido en la Entrega 4 (https://aulas.inf.uva.es/mod/assign/view.php?id=18768)
  - → https://archive.ics.uci.edu/ml/datasets/congressional+voting+records
  - → 435 instancias (50 de cada clase)
  - → 16 atributos (todos nominales)
  - → 1 clase (democrat o republican)
- Thoracic_surgery:Conjunto de datos que hacen referencia a pacientes que se sometieron a interverciones torácicas y síntomas o carácterísticas que estos presentan.
  - → https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data
  - → 470 instancias
  - → 17 atributos
  - → 1 clase
- Energy efficiency: Conjunto de datos que se basa en la evaluación de las necesidades de carga de calefacción y refrigeración de los edificios, es decir su eficiencia energética, en función de los parámetros del edificio.
  - → https://archive.ics.uci.edu/ml/datasets/energy+efficiency
  - → 768 instancias
  - → 8 atributos
  - → 2 clases

2. Ejercicio 1

Conjunto de datos: contact-lenses.arff	Algoritmos (Validación cruzada 10 particiones)				
	J48	OneR	Prism	JRIP	PART
Tasa Error	0,16667	0,291667	0,291667	0,25	0,16667

-Discusión: Podemos ver que los clasificadores que mejores tasas de error nos ofrecen son J48 y PART con un 16.6% de instancias mal clasificadas. Los algoritmos que funcionan peor con este conjunto de datos son OneR y PRISM, lo cuál puede deberse a que son algoritmos simples y a que el conjunto de datos es sumamente reducido.

3. Ejercicio 2

Conjunto de datos: iris.arff	Algoritmos (Validación cruzada 10 particiones)				
	J48	OneR	Prism	JRIP	PART
Tasa Error	0,04	0,08	-	0,046667	0,06

Conjunto de datos: soybean.arff	Algoritmos (Validación cruzada 10 particiones)				
	J48	OneR	Prism	JRIP	PART
Tasa Error	0,084919	0,600293	-	0,077599	0,080527

Conjunto de datos: vote.arff	Algoritmos (Validación cruzada 10 particiones)				
	J48	OneR	Prism	JRIP	PART
Tasa Error	0,036782	0,043678	-	0,045977	0,052874

Conjunto de datos: thoracic_surgery.arff	Algoritmos (Validación cruzada 10 particiones)				
	J48	OneR	Prism	JRIP	PART
Tasa Error	0,155319	0,165957	-	0,153191	0,208511

Conjunto de datos: energyefficiency.arff	Algoritmos (Validación cruzada 10 particiones)				
	J48	OneR	Prism	JRIP	PART
Tasa Error	0,3151	0,8541	-	0,4231	0,3164

-Discusión por algoritmos: Al realizar este ejercicio podemos ver que el algoritmo PRISM no es aplicable a ningún conjunto de datos, ya que este no puede utilizarse si el dataset contiene atributos numéricos o valores desconocidos. En cuanto a los demás algoritmos podemos ver que J48 y JRIP son los que menores tasas de error medio nos ofrece. De nuevo OneR es el algoritmo con el que mayor tasa de error medio obtenemos siendo de un 22.36%.

-Discusión por conjuntos de datos:

Iris: Este conjunto de datos está formado por un número no muy alto de instancias con atributos numéricos y aunque todos los algoritmos nos dan tasas de error bajas, el clasificador que menor tasa de error.

Soybean: Conjunto de datos con gran número de instancias y de atributos que contiene valores desconocidos. La peor tasa de error la obtenemos con OneR, lo cual puede deberse a que hay 19 valores de clase y 35 atributos.

Vote: Para el conjunto de datos de votos para los congresistas de Estados Unidos podemos observar que todos los clasificadores nos ofrecen tasas de error muy bajas próximas a cero, siendo J48 el mejor.

Thoracic_surgery: Para este conjunto de datos la peor tasa de error la obtenemos con el algoritmo PART, lo cual puede deberse a que el conjunto de datos está formado por 683 instancias con 35 atributos tanto numéricos como nominales.

Energy Efficiency: Para este conjunto de datos sigue los patrones anteriormente descritos y la tasa de error aumenta sustancialmente, esto puede ser debido a la discretización llevada a cabo para clasificar.

4. Conclusiones

Tras la realización de los distintos experimentos y el análisis de los resultados contenidos en las tablas de los ejercicios 1 y 2 se pueden apreciar las siguientes peculiaridades: a) el algoritmo J48 es quien mejores resultados obtiene en promedio, b) el algoritmo OneR presenta los peores resultados en promedio con respecto al resto de alternativas, c) en el caso de PRISM y debido a su simplicidad, que no permite la entrada de atributos continuos, tan solo es posible su utilización en el caso del conjunto de datos contact-lenses, en el cual presenta resultados similares a OneR, d) el algoritmo JRIP obtiene tasas de error similares a su homónimo basado en árboles de decisión (J48) pero con tasas de error algo peores en promedio y e) el algoritmo PART obtiene resultados aceptables.

Tras el análisis de los resultados obtenidos, conviene remarcar que la elección entre las distintas estrategias de aprendizaje no es una tarea arbitraria, sino que depende de muchos factores como la estructura del conjunto de datos, la cantidad de instancias de entrenamiento que se posean, las limitaciones computacionales donde se pretenda instaurar el sistema o la tasa de error admisible en la clasificación de resultados.