Reglas clasificación: creación y evaluación de hipótesis con distintos algoritmos

Alejandro Pulido Sánchez

Descripción de los conjuntos de datos

- **Soybean:** formado por **683 instancias** formadas por **35 atributos**, todos ellos de carácter nominal. La clase de destino puede tomar **19 valores distintos**. El conjunto de datos se corresponde con instancias referidas a atributos de plantas y la clase de destinorepresenta el tipo de planta.

- **Vote**: formado por **435 instancias** formadas por **16 atributos**, todos ellos de carácter nominal. La **clase** de destino puede tomar **2 valores distintos**. El conjunto de datos se refiere a resultados de encuestas a ciudadanos estadounidenses para tratar de predecir si votarán al partido demócrata o republicano.

- **Contact-lenses:** Este conjunto de datos está compuesto por **24 instancias** que se dividen en **3 clases diferentes**. Cada instancia representa a un paciente y se le asigna una clase que indica si debe llevar lentes de contacto duras, blandas o ninguna. El conjunto de datos cuenta con **4 atributos nominales** que se utilizan para clasificar a los pacientes en alguna de las 3 clases. Los atributos son la edad del paciente (young, pre-presbyopic, presbyopic), la prescripción de gafas (myope, hypermetrope), la presencia de astigmatismo (no, yes) y la tasa de producción de lágrimas (reduced, normal).

- **Iris:** Este conjunto de datos se compone de **150 instancias**, distribuidas en **3 clases**: iris-setosa, iris-versicolor e iris-virginica. Cada instancia representa una especie de planta Iris y se le asigna una clase correspondiente. El conjunto de datos cuenta con **4 atributos numéricos** que se utilizan para clasificar las instancias en sus respectivas clases. Los atributos son sepallength (longitud del sépalo), sepalwidth (ancho del sépalo), petallength (longitud del pétalo) y petalwidth (ancho del pétalo). Este conjunto de datos se utiliza para el análisis de las características de las plantas Iris.

- **Thoracic_surgery:** El conjunto de datos contiene **470 instancias con 17 atributos**. De estos atributos, hay 16 características médicas de los pacientes, como la edad, el género, la capacidad pulmonar, la gravedad de la enfermedad, el tipo de cirugía, etc. **El último atributo es la clase**, que indica si el paciente murió dentro de un año después de la cirugía o si sobrevivió. El conjunto de **datos es en su mayoría numérico**, aunque **algunos atributos son categóricos**. La mayoría de los valores de los atributos son continuos, pero algunos se presentan en forma de rangos o valores discretos.

- **Breast Cancer Wisconsin (Original):** El conjunto de datos contiene **699 instancias**, cada una de las cuales representa una muestra de células de cáncer de mama. Cada muestra está descrita por **10 atributos numéricos** que corresponden a diferentes características de las células, como el tamaño y la forma de las células. Los atributos incluyen el grosor del tumor, la uniformidad del tamaño de las células, la forma de las células, la adherencia al sustrato y la cantidad de mitosis. Además, cada muestra se etiqueta con **una clase binaria** que indica si la muestra es benigna (no cancerosa) o maligna (cancerosa). La

etiqueta positiva se refiere a células de cáncer maligno, mientras que la etiqueta negativa se refiere a células benignas. El objetivo del análisis de estos datos es construir un modelo que pueda predecir con precisión si una muestra de células es benigna o maligna en función de sus características.

Ejercicio 1

Conjunto de datos: contact-lenses.arff

	Método: 10 XV				
	J48	**OneR**	**Prism**	**JRIP**	**PART**
Tasa error	0.16667	0.291667	0.291667	0.25	0.16667

Los resultados de usar las reglas de clasificación OneR y Prism coinciden con los expuestos en los apuntes de teoría, mientras que PART si que presenta diferencias. Esto se debe a que en los apuntes solo se usan 3 atributos y weka usa todos (5). Los resultados de JRIP no se encuentran en los apuntes.

Ejercicio 2

Conjunto de datos: iris.arff

	Método: 10 XV				
	J48	**OneR**	**Prism**	**JRIP**	**PART**
Tasa error	0.04	0.08	-	0.046667	0.06

No se puede utilizar Prism porque son atributos de tipo real, como se puede ver en lainformación asociada al dataset (https://archive.ics.uci.edu/ml/datasets/iris)

Conjunto de datos: soybean.arff

	Método: 10 XV				
	J48	**OneR**	**Prism**	**JRIP**	**PART**
Tasa error	0.084919	0.600293	-	0.077599	0.080527

No se puede utilizar Prism porque presenta valores desconocidos, como se puede ver en el data set (https://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/soybean.arff)

Conjunto de datos: vote.arff

	Método: 10 XV				
	J48	**OneR**	**Prism**	**JRIP**	**PART**
Tasa error	0.036782	0.043678	-	0.045977	0.052874

No se puede utilizar Prism porque son atributos de tipo categórico y además presenta valores desconocidos, como se puede ver en la información asociada al data set (https://archive.ics.uci.edu/ml/datasets/congressional+voting+records)

Conjunto de datos: thoracic_surgery.arff

	Método: 10 XV				
	J48	**OneR**	**Prism**	**JRIP**	**PART**
Tasa error	0.155319	0.165957	-	0.153191	0.208511

No se puede utilizar Prism porque son atributos de tipo integer y real, como se puede ver en la información asociada al data set (https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data)

Conjunto de datos: Breast Cancer Wisconsin (Original)

	Método: 10 XV				
	J48	**OneR**	**Prism**	**JRIP**	**PART**
Tasa error	0.054363	0.72961	-	0.042918	0.061516

No se puede utilizar Prism porque son atributos de tipo real, como se puede ver en la información asociada al data set (https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28original%29)

Conclusiones

Tras el análisis de los datos obtenidos, es importante destacar que la selección de una estrategia de aprendizaje adecuada no es una decisión aleatoria, sino que depende de varios factores, como la complejidad del conjunto de datos, la cantidad de instancias de entrenamiento disponibles, las limitaciones computacionales en el entorno de implementación del sistema y la tolerancia a la tasa de error en la clasificación de los resultados.