Reglas clasificación: creación y evaluación de hipótesis con distintos algoritmos

Jhon Steeven Cabanilla Alvarado

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 destino representa 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 archivo contiene 24 instancias, 3 clases, una que determina que el paciente debe de estar equipado con lentes de contacto duras, otra que determina que debe de estar equipado con lentes de contacto blandas y por último, la que determina que el paciente no debe estar equipado con ningún tipo de lente. Contiene 4 atributos nominales, que sirven para clasificar al paciente en alguna de las 3 clases, que son la edad del paciente {young, pre-presbyopic, presbyopic}, prescripción de gafas {myope, hypermetrope}, astigmatismo {no,yes} y por último la tasa de producción de lágrimas {reduced, normal}.

- Iris.arff:
Este archivo contiene 150 instancias (50 en cada clase), con 3 clases, iris-setosa, iris-versicolor e iris_virginica. Contiene además 4 atributos numéricos que sirven para clasificar en sus respectivas clases las instancias, el primero sepallength (longitud del sépalo), sepalwidth (ancho del sépalo), petallength (longitud del pétalo) y petalwidth (longitud del pétalo). Este conjunto de datos se corresponde con instancias referidas a atributos de la especie de planta Iris y las clases representan subcategorías de la misma.

- Thoracic_surgery.arff:
Este archivo contiene 470 instancias, y una clase (Risk1Y), así como 16 atributos más la clase que sirven para clasificar las instancias:

1. DGN: Diagnosis - specific combination of ICD-10 codes for primary and secondary as well multiple tumours if any (DGN3,DGN2,DGN4,DGN6,DGN5,DGN8,DGN1)
2. PRE4: Forced vital capacity - FVC (numeric)
3. PRE5: Volume that has been exhaled at the end of the first second of forced expiration - FEV1 (numeric)
4. PRE6: Performance status - Zubrod scale (PRZ2,PRZ1,PRZ0)
5. PRE7: Pain before surgery (T,F)
6. PRE8: Haemoptysis before surgery (T,F)
7. PRE9: Dyspnoea before surgery (T,F)

8. PRE10: Cough before surgery (T,F)
9. PRE11: Weakness before surgery (T,F)
10. PRE14: T in clinical TNM - size of the original tumour, from OC11 (smallest) to OC14 (largest) (OC11,OC14,OC12,OC13)
11. PRE17: Type 2 DM - diabetes mellitus (T,F)
12. PRE19: MI up to 6 months (T,F)
13. PRE25: PAD - peripheral arterial diseases (T,F)
14. PRE30: Smoking (T,F)
15. PRE32: Asthma (T,F)
16. AGE: Age at surgery (numeric)
17. Risk1Y: 1 year survival period - (T)rue value if died (T,F)

- Bank Marketing:
Este archivo contiene 45211 instancias formadas por 17 atributos. Los datos están relacionados con campañas de marketing directo (llamadas telefónicas) de una entidad bancaria portuguesa. El objetivo de la clasificación es predecir si el cliente suscribirá un depósito a plazo.

1 - age (numeric)
2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
5 - default: has credit in default? (categorical: 'no','yes','unknown')
6 - housing: has housing loan? (categorical: 'no','yes','unknown')
7 - loan: has personal loan? (categorical: 'no','yes','unknown')
related with the last contact of the current campaign:
8 - contact: contact communication type (categorical: 'cellular','telephone')
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
other attributes:
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

14 - previous: number of contacts performed before this campaign and for this client (numeric)
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
social and economic context attributes
16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
17 - cons.price.idx: consumer price index - monthly indicator (numeric)
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):
21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

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

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 el link de donde se han obtenido el 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 dataset (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 ademas presenta valores desconocidos, como se puede ver en el link de donde he obtenido el dataset (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 el link de donde he obtenido el dataset (https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data)

Conjunto de datos: bank_marketing.arff

	Método: 10 XV				
	J48	**OneR**	**Prism**	**JRIP**	**PART**
Tasa error	4.38	0.885625	-	202203.311	4.926

No se puede utilizar Prism porque son atributos de tipo numérico, como se puede ver en el link de donde he obtenido el dataset (https://archive.ics.uci.edu/ml/datasets/bank+marketing)

Conclusiones

Tras el análisis de los datos 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.