ENTREGA 5

PRÁCTICA REGLAS DE CLASIFICACIÓN: CREACIÓN Y EVALUACIÓN DE HIPÓTESIS CON DISTINTOS ALGORITMOS (Y COMPARACIÓN DE ÁRBOLES).

Patricia Aguado Labrador

Técnicas de Aprendizaje Automático
Patricia Aguado Labrador
2019/2020

DESCRIPCIÓN DE LOS CONJUNTOS DE DATOS QUE SE UTILIZARÁN

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
5 atributos (4 + 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
5 atributos (4 + clase)
3 clases

Soybean: explicada en la práctica 4.

https://archive.ics.uci.edu/ml/datasets/Soybean+(Large)
683 instancias
36 atributos (35 + clase)
19 clases

Vote: explicada en la práctica 4.

https://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records
435 instancias
17 atributos (16 + clase)
2 clases

Thoracic_surgery: explicada en la práctica 4.

https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data
470 instancias
17 atributos (16 + clase)
2 clases

Biodeg: dataset fromado por valores experimentales de biodegradación de 1055 productos químicos. Los 41 atributos se utilizan para desarrollar modelos de relaciones cuantitativas de actividad de estructura, con el fin de discriminar moléculas biodegradables listas y no listas.

https://archive.ics.uci.edu/ml/datasets/QSAR+biodegradation
1055 instancias
42 atributos(41 + clase)
2 clases

EJERCICIO 1

Reglas con PRISM y conjunto de datos: contact-lenses

If astigmatism = no
and tear-prod-rate = normal
and spectacle-prescrip = hypermetrope then soft

If astigmatism = no
and tear-prod-rate = normal
and age = young then soft

If age = pre-presbyopic
and astigmatism = no
and tear-prod-rate = normal then soft

If astigmatism = yes
and tear-prod-rate = normal
and spectacle-prescrip = myope then hard

If age = young
and astigmatism = yes
and tear-prod-rate = normal then hard

If tear-prod-rate = reduced then none

If age = presbyopic
and tear-prod-rate = normal
and spectacle-prescrip = myope
and astigmatism = no then none

If spectacle-prescrip = hypermetrope
and astigmatism = yes
and age = pre-presbyopic then none

If age = presbyopic
and spectacle-prescrip = hypermetrope
and astigmatism = yes then none

Reglas con OneR y conjunto de datos: contact-lenses

tear-prod-rate:
reduced -> none
normal -> soft

Podemos ver, comparando con las reglas que están reflejadas en los apuntes de clase, que son exactamente iguales a las producidas con Weka..

Conjunto de datos: **contact-lenses**	
Algoritmo	Tasa de error
J48	0.166
OneR	0.292
PRISM	0.292
JRIP	0.25
PART	0.166

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 cual puede deberse a que son algoritmos simples y a que el conjunto de datos es sumamente reducido.

EJERCICIO 2

Tasas de error					
Algoritmo	Conjuntos de datos				
	Iris	Soybean	Vote	Thoracic_surgery	Biodeg
J48	0.04	0.085	0.037	0.155	0.176
OneR	0.08	0.6	0.044	0.166	0.228
PRISM	x	x	x	x	x
JRIP	0.053	0.078	0.046	0.153	0.176
PART	0.06	0.081	0.053	0.208	0.148

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 nos da es J48.

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.

Biodeg:

Para este conjunto de datos formado por 1055 instancias con 42 atributos, el algoritmo que peor funciona es OneR, lo cual puede ser debido a la cantidad de pares atributo-valor que se evaluan en la construcción de la regla.

CONCLUSIONES

En esta práctica he aprendido como se comportan los diferentes algoritmos basados en la construcción de reglas y de árboles, así como ver las características de cada uno.