OCR-Weka: Reglas de clasificación – PART

NOMBRE: Frédéric Roux **Fecha: 13/12/2018**

En la carpeta *classifiers/rules* de la pestaña *Classify* de Weka puedes encontrar diferentes clasificadores basados en reglas. La inducción de reglas se puede lograr fundamentalmente mediante dos caminos: generando un árbol de decisión y extrayendo de él las reglas, como puede hacer el sistema C4.5, o bien mediante una estrategia de *covering*, consistente en tener en cuenta cada vez una clase y buscar las reglas necesarias para cubrir [*cover*] todos los ejemplos de esa clase; cuando se obtiene una regla se eliminan todos los ejemplos que cubre y se continúa buscando más reglas hasta que no haya más ejemplos de la clase.

(a) Por ejemplo, el algoritmo **PART** (*obtaining rules from PARTial decision trees*) adopta la estrategia del covering y la generación de reglas a partir de árboles de decisión. Para crear una regla utiliza un árbol de decisión podado, se obtiene la hoja que clasifique el mayor número de ejemplos, que se transforma en regla, se eliminan los ejemplos que dicha regla cubre y se continúa generando reglas hasta que no queden ejemplos por clasificar.

El proceso de elección del mejor atributo se hace como en el sistema C4.5, por lo que los parámetros del algoritmo PART son un subconjunto de los ofrecidos por J48, que implementa el sistema C4.5: *minNumObj, confidenceFactor, unpruned*, etc.

Prueba este algoritmo sobre el ejemplo de OCR utilizando los valores por defecto para los parámetros. Incluye en este documento las reglas obtenidas. ¿Qué tasa de acierto se obtiene en este caso para el conjunto de test (tras entrenar el clasificador con el conjunto de entrenamiento)? ¿Se parecen las reglas a las obtenidas con el clasificador J48 (C4.5)?

La tasa de acierto para el conjunto test con el algoritmo PART es de 59%.

PART decision list
---

14 <= 1 AND
5 <= 30: 6 (11.0/1.0)

31 <= 0 AND
22 <= 5 AND
14 <= 75: 7 (10.0)

31 <= 0 AND
6 <= 55: 4 (11.0/1.0)

14 > 10 AND
27 <= 6 AND
31 > 3: 3 (11.0/1.0)

14 > 10 AND

27 <= 4: 9 (10.0)

33 <= 92 AND
20 <= 47 AND
14 > 10 AND
10 > 24: 8 (10.0)

28 > 81: 1 (10.0)

15 <= 16 AND
5 <= 33: 2 (9.0)

15 <= 12: 5 (9.0)

: 0 (9.0)

Number of Rules : 10

Existen algunas similitudes entre las reglas obtenidas con el algoritmo PART y el arbol de decision J48, pero no son iguales. Por ejemplo la clase 6 depende de las variables 14 (14<=1) y 5 (5<=30) para ambos algoritmos, pero la clase 5 depende de distintas variables: 15 (<=16) y 5 (<=33) en el caso de PART, y 14 (<=1) y 5(>30) en el caso del J48.

(b) En la misma carpeta existen otros algoritmos basados en reglas. Prueba alguno de ellos, indicando los parámetros utilizados y el significado de los mismos, así como los resultados de clasificación obtenidos. Recuerda entrenar los clasificadores con la muestra de entrenamiento.

Algoritmo: JRIP

Parámetros:
*seed – semilla para randomizacion
*numDecimalPlaces – precisión decimal para números
* batchSize – el numero de casos usado para el entrenamiento.
* folds – el numero de casos usados para el pruning. El resto de los casos estan usados para determinar las reglas.
* minNo – el peso mínimo que tiene una variable predictiva en una regla
*optimizations – el numero de pasos de optimización
*doNotCheckCapabilities – si esta opción esta activada, las capacidad/validez del clasificador no esta comprobada
*checkErrorRate – decide si o no el criterio de tasa de error >= 0.5 esta incluido en el criterio de stop (se para la optimización cuando la tasa de error >=0.5)
*usePruning – decide si o no el árbol esta podado

Tasa de asierto: 44%