Ejercicios Prácticos Jesús García Herrero	

Análisis de datos de imágenes con Weka

Tenemos datos de segmentación de imágenes. Cada instancia de los datos proporciona información de una región de 3x3 píxeles (parche). Por cada instancia, tenemos los siguientes atributos:

1. **TIPO_IMAGEN:** Clase a la que pertenece el parche (brickface, sky, foliage, cement, window, path, grass).
2. **REGION-CENTROID-COL:** Número de la columna que ocupa el píxel central del parche en la imagen original.
3. **REGION-CENTROID-ROW:** Número de la fila que ocupa el píxel central del parche en la imagen original.
4. **SHORT-LINE-DENSITY-5:** Los resultados de un algoritmo de extracción que cuenta cuantas líneas de longitud 5 (en cualquier orientación) con poco contraste (menos o igual que 5) van en la región.
5. **SHORT-LINE-DENSITY-2:** Igual que el anterior pero cuenta líneas con un contraste mayor que 5
6. **VEDGE-MEAN:** Media de la medida del contraste de los píxeles verticales adyacentes de la región (existen 6).
7. **VEDGE-SD:** Desviación de la medida del contraste de los píxeles verticales adyacentes de la región (existen 6).
8. **HEDGE-MEAN:** Media de la medida del contraste de los píxeles horizontales adyacentes de la región (existen 6).
9. **HEDGE-SD:** Desviación de la medida del contraste de los píxeles horizontales adyacentes de la región (existen 6).
10. **INTENSITY-MEAN:** Media de la intensidad de la región (R+G+B)/3
11. **RAWRED-MEAN:** La media de color rojo en la región
12. **RAWBLUE-MEAN:** La media de color azul en la región
13. **RAWGREEN-MEAN:** La media de color verde en la región
14. **EXRED-MEAN:** Medida de exceso de rojo. (2R – (G+B))
15. **EXBLUE-MEAN:** Medida de exceso de azul. (2B – (G+R))
16. **EXGREEN-MEAN:** Medida de exceso de verde.(2G – (B+R))
17. **VALUE-MEAN:** Valor "v" medio del modelo hsv de la región
18. **SATURATION-MEAN:** Saturación "s" media del modelo hsv de la región
19. **HUE-MEAN:** Tonalidad "h" media del modelo hsv de la región**.**

Para la realización de la práctica trabajaremos con los siguientes ficheros:

- segmentationDATA.arff: Conjunto de datos de entrenamiento
- segmentationTEST.arff: Conjunto de datos de test
- segmentationReducidosAdaptados.xls: Adaptación de los datos del conjunto de entrenamiento para "discretizarlos" (y conversión del tipo de fichero a EXCEL)
- segmentationReducidosAdaptados.csv: Extracción de los datos del fichero EXCEL a formato tecto separado por *punto_y_coma*
- segmentationReducidosAdaptados.arff: conversión de los datos del fichero *.csv para su tratamiento en WEKA

Se pide:

1. Eliminar las colunas REGION-CENTROID-COL, REGION-CENTROID-ROW y REGION-PIXEL-COUNT usando los filtros de Weka.
2. Generar árbol C4.5 (J48 en Weka) para determinar el atributo TYPE a partir del resto usando un 70% de los datos para construir el árbol y un 30% para validar los resultados. ¿Qué resultados obtiene? ¿Como son comparados con los obtenidos en la práctica anterior?
3. Generar reglas usando los algoritmos PRISM y PART usando la misma metodología que en el inciso anterior y comparar con las generadas anteriormente.

Parte de filtrado:

4. Usar el tab “Select Atributes” para determinar que columnas del conjunto de datos son más relevantes para clasificar correctamente el TYPE. Usar para evaluar los atributos un clasificador que aplique el J48 y como método de búsqueda un algoritmo genético.
5. ¿Qué variables son más significativas para predecir el TYPE?
6. Compare los resultados de aplicar un J48 con la metodología del ejercicio anterior pero que use sólo las variables que encontró como importantes en el inciso anterior.