Bases de Datos Masivas (11088)
Departamento de Ciencias Básicas

TRABAJO PRÁCTICO III: Minería de datos

PARTE 01: Arboles de decisión (J48)

Introducción:

En este trabajo se implementa el primero de una serie de algoritmos que se presentarán durante la materia para realizar Minería de datos: los árboles de decisión J48.

En primer lugar, se utilizarán las nociones de entropía y ganancia de información introducidas en clase a efectos de generar un árbol de decisión a partir de un dataset.

Luego, se utilizará el software **Weka** con el objetivo de resolver problemas de la disciplina, los cuales son una combinación ejercicios clásicos de minería de datos complementados con ejercicios propuestos por el equipo docente.

Consignas:

1. A partir del dataset presentado a continuación, y teniendo en cuenta las fórmulas de entropía y ganancia de información calcule y diagrame el árbol de decisión que le permita decidir si comer asado o no en función del clima:

PRONÓSTICO	TEMPERATURA	HUMEDAD	VIENTO	ASADO
Soleado	Calor	alta	leve	no
Soleado	Calor	alta	fuerte	no
Nublado	Calor	alta	leve	si
Lluvioso	templado	alta	leve	si
Lluvioso	Frío	normal	leve	si
Lluvioso	Frío	normal	fuerte	no
Nublado	Frío	normal	fuerte	si
Soleado	templado	alta	leve	no
Soleado	Frío	normal	leve	si
Lluvioso	templado	normal	leve	si
Soleado	templado	normal	fuerte	si

Nublado	templado	alta	fuerte	si
Nublado	Calor	normal	leve	si
Lluvioso	templado	alta	fuerte	no

2. Genere el archivo .arff de acuerdo al "Anexo I – Formato .arff para Weka" y el árbol de decisión mediante esa herramienta. Compare el árbol resultante con el del punto 1:
    a. ¿Encuentra diferencias entre ambos?
    b. ¿Cuáles instancias fueron clasificadas de manera incorrecta en cada uno? ¿A qué se debe?
3. Desarrolle el algoritmo de árbol de decisión a partir de un lenguaje de programación.
4. Cargue en Weka la base de datos zoo.arff:
    a. Explore las características de los datos, instancias, cantidad de atributos, su significado y el tipo de datos de cada uno.
    b. Genere el árbol de decisión J48 que permita inferir el tipo de animal en función de sus características. Explique someramente que resultado se obtiene en términos del conocimiento aportado por el árbol.
        - ¿Varía ese resultado si se elimina el atributo "animal"? ¿Por qué?
        - Cuantos niveles posee el árbol generado? ¿Qué atributos debemos modificar si deseamos realizar una poda del mismo? Modifique esos atributos para que el árbol generado conste de 4 niveles. ¿Afecta el *accuracy* de la clasificación esta modificación?
    c. Si hubiera querido generar un árbol ID3, ¿Hubiera podido? ¿Por qué?
5. Se provee la base de datos de los pasajeros del famoso barco que se hundiera en su viaje inaugural (archivo titanic-en.arff) con los siguientes atributos y valores posibles:

```
- @ATTRIBUTE class {"1st","2nd","3rd","crew"}
- @ATTRIBUTE age {"adult","child"}
- @ATTRIBUTE sex {"male","female"}
- @ATTRIBUTE survived {"yes","no"}
```

Genere el árbol de decisión J48, explore y documente la solución dada.

6. Un Banco de Portugal realizó una campaña de marketing en busca de clientes de plazos fijos basada en llamados telefónicos. Se provee el dataset real (bank-full.csv) con más 45000 instancias y el detalle (bank-names.txt) de los datos registrados de cada una de las personas contactadas por la entidad bancaria.
    a. Realice las tareas necesarias para poder procesar el dataset en Weka.
    b. Luego, genere el árbol de decisión, y optimice los resultados, con el objetivo de explicar cuáles son las características más importantes que permiten identificar a una persona que accederá o no al plazo fijo.
    c. Documente los resultados.
7. **Medidas de evaluación para técnicas de clasificación.** En función de la clasificación realizada en 2), 4), 5) y 6) complete las siguientes actividades:
    a. **Curva ROC.**
        1. Genere la matriz de confusión de las clasificaciones y grafique la curva ROC.
        2. ¿Qué información brinda la técnica del punto anterior?
        3. ¿Qué permite concluir con respecto a las clasificaciones?
    b. **Accuracy.**
        1. Ahora, calcule el accuracy de los modelos.
        2. ¿Cómo se interpreta la métrica anterior?
        3. ¿Qué aporta el accuracy?
    c. **Recall/Precision.**

1. Calcule las métricas recall y precisión para los modelos.
2. ¿Cuál es la diferencia entre ambas?
3. ¿Qué aspectos aborda cada una?

d. **F-score.**

1. Calcule la métrica F-score en función de la matriz de confusión resultante.
2. ¿Qué haría en caso de querer dar mayor importancia a la precisión del modelo? ¿Y en caso de querer ponderar la exhaustividad?
3. Compare y documente los resultados.

e. **Kappa.** Calcule el Coeficiente de Kappa. ¿En qué casos resulta conveniente?

8. Explore para dos de los modelos generados antes los diferentes métodos de validación disponibles en Weka. ¿Varía la performance del árbol de acuerdo al método? ¿A qué se debe? Explique someramente cada uno.
9. Guarde los archivos resultantes de las actividades prácticas en una carpeta denominada tp0301-<legajo> que a su vez tenga un directorio por cada uno de los puntos de este trabajo, comprima la carpeta y envíelo al equipo docente.

Referencias sugeridas:

Data Mining: Practical Machine Learning Tools and Techniques http://www.cs.waikato.ac.nz/ml/weka/book.html

Machine Learning, Chapter 3. Tom M. Mitchell, McGraw Hill, 1997.

Introducción a la Minería de Datos. Hérnandez Orallo & otro. Prentice Hall. 2008.

ANEXO I:

FORMATO ARFF PARA WEKA

Si bien la aplicación Weka permite el ingreso de datos desde diferentes tipos de fuentes de datos, incluso a través de conexiones a Bases de Datos, muestra su mejor performance a partir de archivos ARFF (archivos planos organizados en filas y columnas), los cuales tienen la siguiente estructura:

@relation <nombre_db>

@attribute <nombre_atributo1> <tipo_dato>

@attribute <nombre_atributo2> <tipo_dato>

…

@attribute <nombre_atributoN> <tipo_dato>

@data

<valor_dato1_atributo1>,<valor_dato1_atributo2>,…,<valor_dato1_atributoN>

<valor_dato2_atributo1>,<valor_dato2_atributo2>,…,<valor_dato2_atributoN>

…

<valor_datoN_atributo1>,<valor_datoN_atributo2>,…,<valor_datoN_atributoN>

Algunas aclaraciones al respecto:

- **@relation**: Se define arbitrariamente el nombre para el set de datos,
- **@attribute**:
    - En cuanto al nombre, se define arbitrariamente el nombre para el atributo,
    - En cuanto al tipo de dato, algunos de los más importantes se enumeran a continuación:
        - **numeric,**
        - **string**,
        - **date <formato_fecha>:** donde el formato de fecha se especifica, por ejemplo, de la siguiente manera: "dd-mm-yyyy HH:mm:ss".
        - **Discretos:** donde los posibles valores se definen entre llaves (‘{‘, ‘}’) y separados por comas. Por ejemplo: @attribute sexo {M, F}
- **@data:** En la línea siguiente a esta etiqueta se incorpora el set de datos.