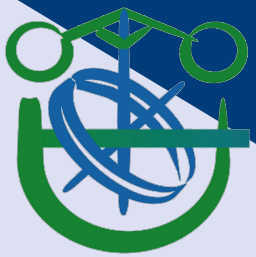

Escuela Politécnica Superior (EPS)

UNIVERSIDAD D CÓRDOBA
Departamento de Informática y Análisis Numérico

Práctica 1 - 2 Filtros en Weka

Juan Carlos Fernández Caballero (jfcaballero@uco.es)

Introducción al Aprendizaje Automático (IAA)
3º de Grado de Ingeniería Informática
Especialidad en Computación
Curso 2019-2020

GRUPO DE INVESTIGACIÓN AYRNA
APRENDIZAJE **Y** **R**EDES **N**EURONALES **A**RTIFICIALES
uco.es/ayrna

Índice de contenidos

Filtros

Ejercicios

Entregables

Bibliografía

Filtros supervisados vs No supervisados

- De forma general, **los filtros persiguen lo siguiente**: Transformar los datos para un mejor aprendizaje y adaptación de los modelos $\longrightarrow$ **Preprocesado**.
- Ejemplos de algunas **operaciones sobre los datos** son: discretizar, normalizar, eliminar valores, sustituir valores.
- Los hay de 2 tipos, supervisados y no supervisados [1]:

Filtros supervisados

- **Tienen en cuenta el ultimo atributo del dataset** a la hora de hacer un tratamiento sobre los datos.
- En clasificación será la **clase asignada a un patrón**. En caso de regresión, el **valor de salida** a predecir.

Filtros supervisados vs No supervisados

Filtros no supervisados

- **No tienen en cuenta el ultimo atributo** del dataset a la hora de hacer un tratamiento sobre los datos.
- Por defecto toman el último atributo como clase o valor numérico de salida para regresión (*ignore class* $\rightarrow$ *false*), **aplicándose el filtro a todos los patrones y atributos (menos al último atributo)**.
- Si queremos cargar una serie de datos a los que **aplicar filtros en su totalidad**, indicar en el filtro *ignore class* $\rightarrow$ *true*.

Filtros para transformación de los datos

Filtros para **transformación de los datos:**

Objetivos

- Reemplazar datos perdidos.
- Reducir el tamaño de los datos para su procesamiento.
- Eliminar datos repetitivos o que no aportan información.

Algunos filtros para transformación de los datos:

- **filters/unsupervised/attribute/Normalize**
- **filters/unsupervised/attribute/ReplaceMissingValues**
- **filters/unsupervised/attribute/NominalToBinary**. Se explica a continuación...

Filtros para transformación de los datos

filters/unsupervised/attribute/NominalToBinary

- La opción *BinaryAttributesNominal=True* hace que los atributos binarios resultantes pongan un valor '0' como ausencia y no como valor numérico.
- IMPORTANTE: Si el atributo nominal solo tiene 2 valores, pasa a ser numérico con valores 0 y 1, a no ser que se configure la opción *transformAllValues=True*.

filters/supervised/attribute/NominalToBinary

- @attribute miAtributo {1,2,3}
- Para el caso de que los valores nominales sean un conjunto de números y no cadenas, podría usarse también el filtro supervisado.
- Este filtro transformaría los "N" valores numéricos del atributos nominal a "N-1" valores (consultar ayuda de Weka para ver cómo).

Filtros para selección de características (atributos)

Filtros para **selección de características (atributos):**

Objetivos

- Reducir el coste computacional asociado al aprendizaje, eliminando atributos irrelevantes o redundantes.
- Mejorar la calidad del modelo y expresarlo de forma más comprensible, eliminando atributos que son perjudiciales para el aprendizaje.
- A esto se le llama reducción de dimensionalidad.

Algunos filtros:

- **filters/unsupervised/attributes/RemoveUseless**

Filtros para selección de patrones (instancias)

Filtros para **selección de patrones (instancias):**

Objetivos

- Reducir el número de patrones de las clases más numerosas.
- Incrementar el número de patrones de las clases minoritarias, introduciendo patrones sintéticos.
- Eliminar instancias problemáticas.
- Extraer instancias representativas de las clases.

Algunos filtros:

- **filters/supervised/instance/SpreadSubsample**
- **filters/supervised/instance/ClassBalancer**
- **filters/supervised/instance/Resample**
- **filters/unsupervised/instance/Resample** ¿en que se diferenciará del *supervised*?
- **filters/unsupervised/instance/RemovePercentage**

Ejercicios

1. Cargue la base de datos Iris (disponible en Moodle). Observe los atributos.
   - **1.1** ¿Cuántos atributos caracterizan los datos de esta base de datos?
   - **1.2** ¿Se trata de regresión o clasificación?
   - **1.3** ¿Cuál es el rango de valores del atributo petalwidth?¿Y su media? ¿y su desviación típica?
   - **1.4** Utilizando el entorno *Weka Explorer* $\rightarrow$ *Visualize*, determinar que atributo permite discriminar linealmente entre la clase iris-setosa y las otras dos clases.
   - **1.5** ¿Es posible separar linealmente la clase iris-versicolor de la clase iris-virginica?
   - **1.6** ¿Con qué dos atributos te quedarías para discriminar entre las tres clases del problema?
   - **1.7** ¿Que diferencia hay entre instancias *Distinct* y *Unique*? Fabríquese una base de datos pequeña propia para poner un ejemplo.

Ejercicios

2. Cargue la base de datos *audiology* (disponible en Moodle).
   - **2.1** Aplique el filtro **filters/unsupervised/attribute/NominalToBinary** y describa como quedan ahora los atributos.
   - **2.2** ¿Podría saber con antelación el número de atributos finales al aplicar este filtro?
   - **2.3** ¿Que ha pasado con algunos atributos nominales?
3. Particione una base de datos usando **filters/supervised/instance/StratifiedRemoveFolds**
   - **3.1** Divida el dataset en *train* y *test* mediante un *3-fold*.
   - **3.2** Divida el dataset en *train* y *test* mediante un *3-holdOut* con un 75 % train y 25 % test.

Entregables

1. Elija 3 **filtros No Supervisados** de los que aparecen listados, expliquelos y describa cómo quedan los datos antes y después al aplicarlos sobre una o varias bases de datos.
   - Consulte el *UCI Machine Learning Repository* para una descripción de la base de datos y la transformación a .arff
   - Si no puede aplicar un filtro elegido en ninguna base de datos describa por qué, y construyase una base de datos ficticia y pequeña donde si pueda aplicarlo.
   - Use capturas de pantalla, salidas de Weka y todo lo que considere necesario para sus ejercicios.
   - La puntuación variará en función de la argumentación y dificultad de los filtros elegidos.
     1.1 **filters/unsupervised/attribute/Normalize**
     1.2 **filters/unsupervised/attribute/ReplaceMissingValues**
     1.3 **filters/unsupervised/attributes/NominalToBinary**
     1.4 **filters/unsupervised/intance/RemoveDuplicates**
     1.5 **filters/unsupervised/instance/Resample**.
     1.6 **filters/unsupervised/attribute/Remove**
     1.7 **filters/unsupervised/attributes/RemoveUseless**

Entregables

2. Elija 3 **filtros Supervisados** de los que aparecen listados, expliquelos y describa cómo quedan los datos antes y después al aplicarlos sobre una o varias bases de datos.

- Consulte el *UCI Machine Learning Repository* para una descripción de la base de datos y la transformación a .arff
- Si no puede aplicar un filtro elegido en ninguna base de datos describa por qué, y construyase una base de datos ficticia y pequeña donde si pueda aplicarlo.
- Use capturas de pantalla, salidas de Weka y todo lo que considere necesario para sus ejercicios.
- La puntuación variará en función de la argumentación y dificultad de los filtros elegidos.

2.1 filters/supervised/attribute/Discretize
2.2 filters/supervised/attribute/NominalToBinary
2.3 filters/supervised/instance/SpreadSubsample
2.4 filters/supervised/instance/ClassBalancer
2.5 filters/supervised/instance/Resample

Bibliografía adicional a la de la asignatura y al material de Moodle

 Weka 3: Data Mining Software in Java, 2019. https://www.cs.waikato.ac.nz/ml/weka.

¿Preguntas?