MÁSTER OFICIAL DE LA UNIVERSIDAD DE GRANADA

"SOFT COMPUTING Y SISTEMAS INTELIGENTES"

M1. FUNDAMENTOS DE MINERÍA DE DATOS

J.L. CUBERO, F. BERZAL, F. HERRERA

Dpto. Ciencias de la Computación e I.A.

Universidad de Granada

18071 – ESPAÑA

DECSAI

Universidad de Granada

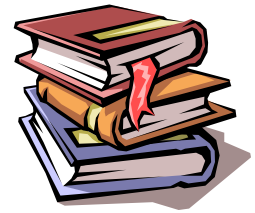
Contenido del Curso

Estudio práctico de las técnicas de preparación de datos:

Limpieza
Transformación
Reducción de datos

http://www.cs.waikato.ac.nz/ml/weka/

Bibliografía



Bibliografía – Minería de Datos con WEKA

Ian H. Witten & Eibe Frank
DATA MINING
Practical Machine Learning Tools and Techniques
Second Edition
Morgan Kaufmann
2005

Preprocesamiento con WEKA

- **Introducción a WEKA**
- **Técnicas básicas de preprocesamiento en WEKA**
- **Reducción de datos en WEKA**
  - **Discretización**
  - **Selección de Características**
  - **Selección de Instancias**
- **Ejemplos Prácticos y Ejercicios**

Introducción a WEKA

WEKA es una recopilación de algoritmos para aprendizaje automático y herramientas de preprocesamiento de datos.

Además proporciona soporte para todo el proceso experimental: evaluación, preparación y visualización de datos y resultados.

WEKA contiene métodos de clasificación, regresión, clustering y reglas de asociación

Introducción a WEKA

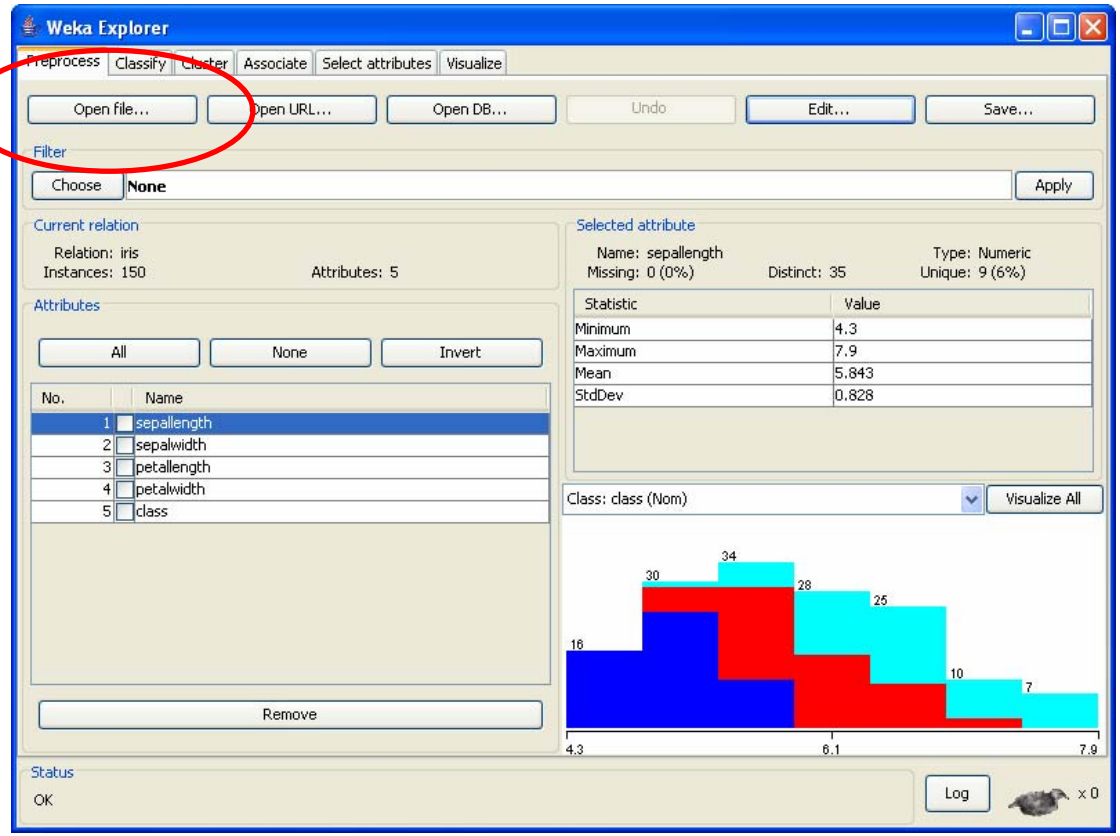
Maneras de Interacción:

- **Explorer:** Se utiliza para ejecutar y comparar resultados sobre un único conjunto de datos.
- **Experimenter:** Para construir una experimentación completa y almacenar resultados.
- **Knowledge Flow:** IDEM a experimenter pero representa el experimento con un grafo dirigido.

Introducción a WEKA

Explorer

- Cargar Datos
- Soporta ficheros en formato ARFF, CSV, Excel y conexión jdbc con BDs.



Introducción a WEKA

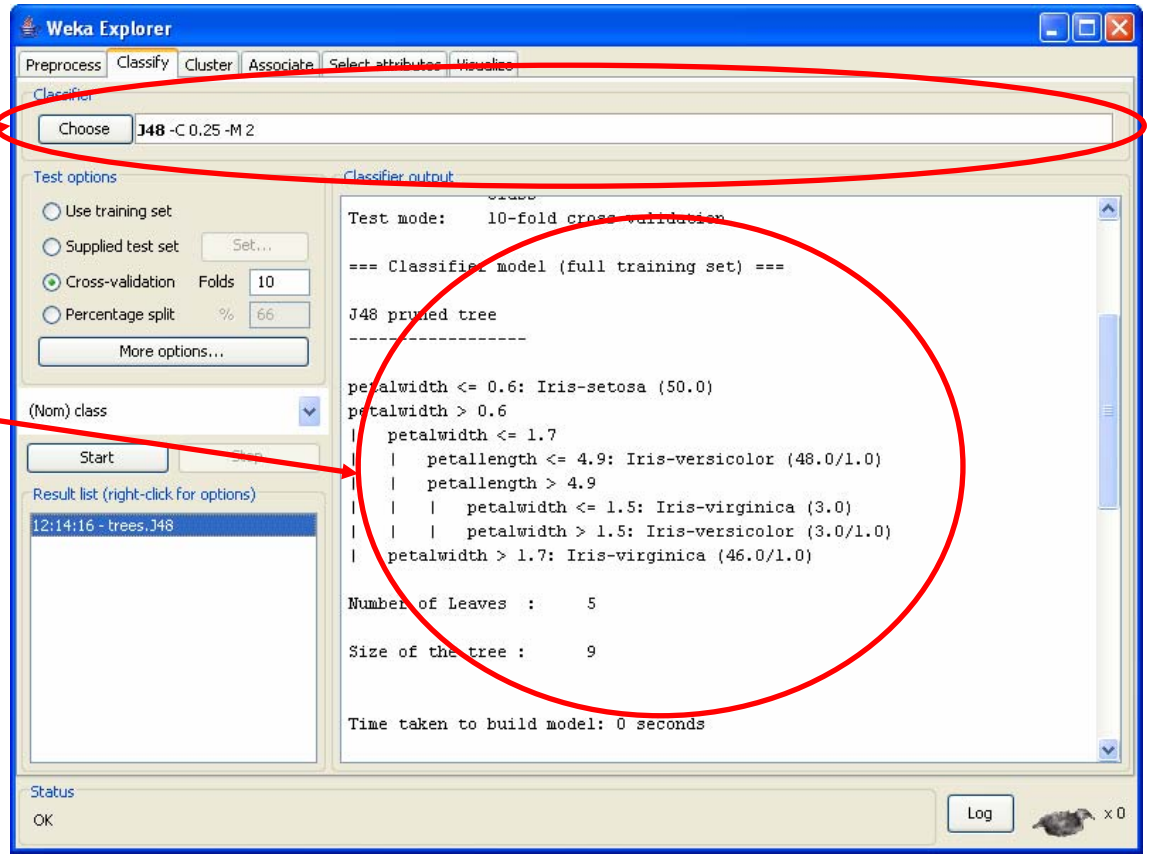
Formato ARFF

```
@RELATION iris

@ATTRIBUTE sepallength  REAL
@ATTRIBUTE sepalwidth   REAL
@ATTRIBUTE petallength  REAL
@ATTRIBUTE petalwidth   REAL
@ATTRIBUTE class        {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
5.2,2.7,3.9,1.4,Iris-versicolor
5.0,2.0,3.5,1.0,Iris-versicolor
7.2,3.6,6.1,2.5,Iris-virginica
6.5,3.2,5.1,2.0,Iris-virginica
```

Introducción a WEKA

Explorer

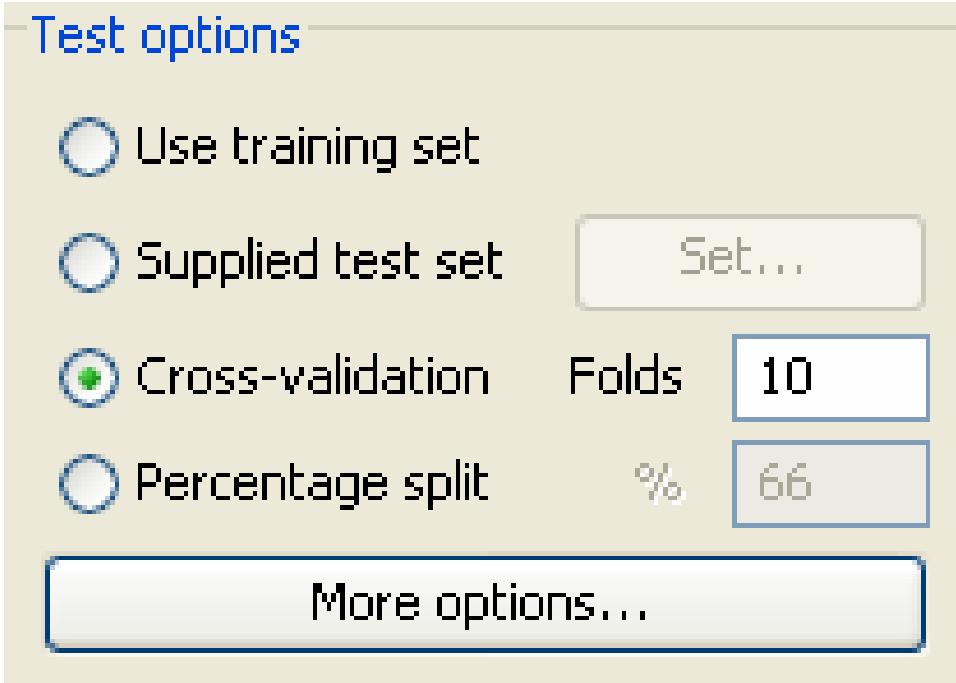
- Construir un árbol de decisión
- Visualizar resultados de la ejecución



Introducción a WEKA

Explorer

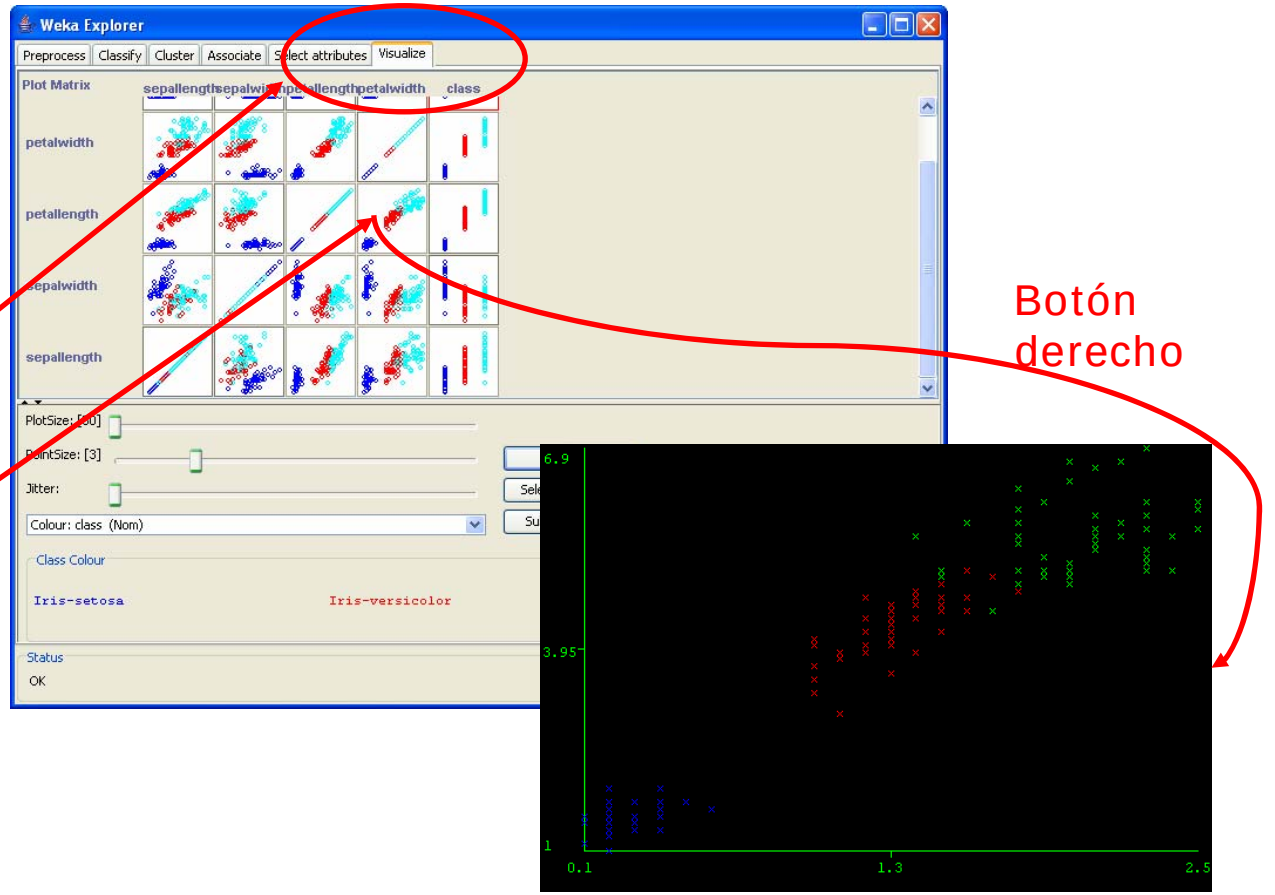
- Formas de comprobar resultados (test):
  - Solo entrenar
  - Usar un conjunto de test específico
  - k-cfv
  - Particion determinada



Introducción a WEKA

Explorer

- Visualización de Datos
  - Atributos vs. Atributos
  - Detalle

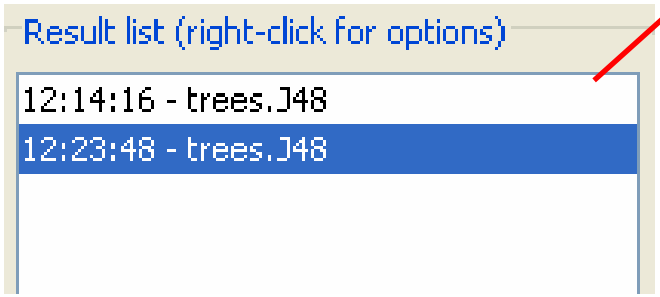


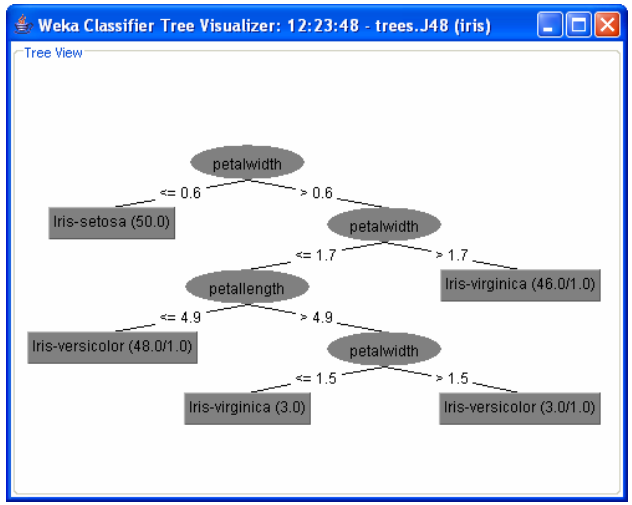
Introducción a WEKA

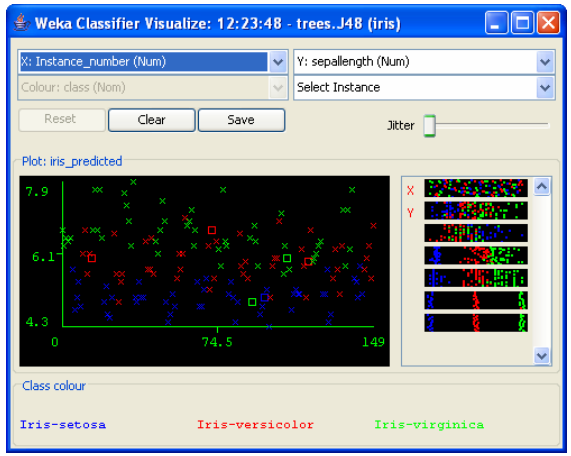
Explorer

- Visualización de Resultados y modelos

Botón derecho







Introducción a WEKA

Explorando Explorer (algoritmos más conocidos)

- Clasificación

BayesNet: Aprende redes Bayesianas

NaiveBayes: Clasificador discriminador bayes standard

Id3: Divide y Vencerás básico para árboles de decisión

J48: C4.5

RandomForest: Construye un Bosque Aleatorio

JRip: Algoritmo RIPPER

M5Rules: Construye reglas M5 desde árboles

LinearRegression: Regresión linear standard

MultilayerPerceptron: Red neuronal de retropropagación

RBFNetwork: Red de función radio base

SMO: Clasificación basada en vectores de soporte

Ibk: k vecinos más cercano

LWL: Aprendizaje basado en pesos locales

Introducción a WEKA

Explorando Explorer (algoritmos más conocidos)

- Clustering
  - CobWeb: Algoritmo CobWeb
  - SimpleKMeans: Algoritmo k-Medias
- Asociación
  - Apriori: Algoritmo Apriori
  - PredictiveApriori: A priori con orden según acierto predictivo.

Introducción a WEKA

Knowledge Flow

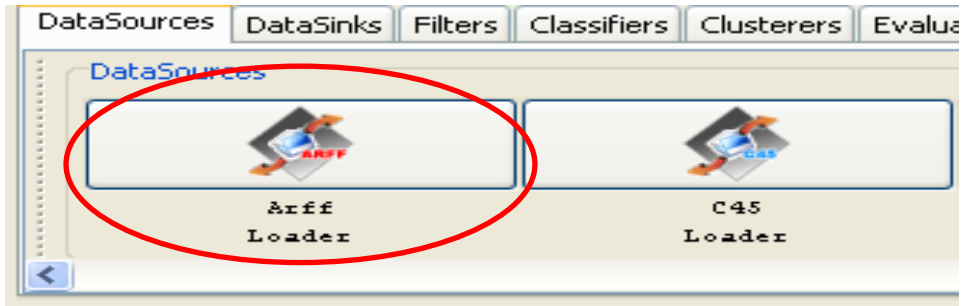
- Proporciona una alternativa a *Explorer* para aquellos que piensan en términos de como los datos fluyen a través del sistema.
- Permite crear configuraciones imposibles por *Explorer*.

Introducción a WEKA

Knowledge Flow: Ejemplo paso a paso (1)

- Objetivo: Cargar un fichero ARFF y llevar a cabo una validación cruzada con C4.5

- Pulsar *Arff Loader* dentro de la pestaña *Data Sources.* Pinchar en el panel de Dibujo.



Introducción a WEKA

Knowledge Flow: Ejemplo paso a paso (2)

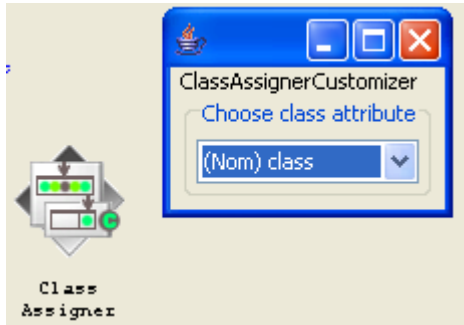
- Con el botón derecho en el nodo *Arff Loader*, ir a *Configure*. Por ejemplo, trabajaremos con *iris.arff.*

- Para indicar el atributo que será la clase, introducir un nodo *Class Asigner* (Etiqueta *Evaluation*).

Introducción a WEKA

Knowledge Flow: Ejemplo paso a paso (3)

- Para conectar ambos nodos, pulsar con el botón derecho sobre *Arff Loader* y seleccionar la opción *Conections->dataset.*
- Configurar el *Class Asigner* para tomar el atributo *Nom* como clase.



Introducción a WEKA

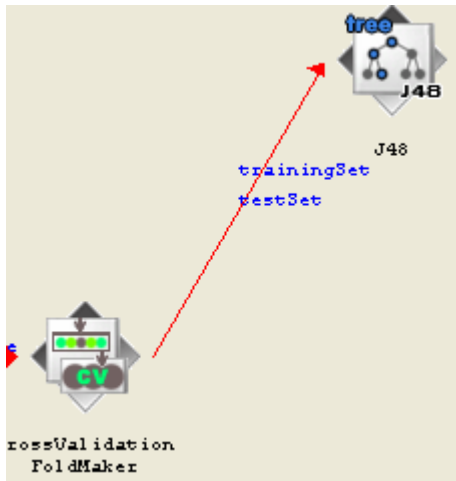
Knowledge Flow: Ejemplo paso a paso (4)

- Hacer la misma operación para un *CrossValidation Foldmaker.*
- Configurarlo con 10 folds y poner una semilla cualquiera.

Introducción a WEKA

Knowledge Flow: Ejemplo paso a paso (5)

- Introducir un nodo *J48* (en pestaña *classifiers*).
- Conectarlo con *CrossValidation Foldmaker* mediante la conexión *trainingSet* y *testSet*.



Introducción a WEKA

Knowledge Flow: Ejemplo paso a paso (6)

- Introducir un nodo *Classifier PerformanceEvaluator* (en pestaña *evaluation*).
- Conectarlo con *J48* mediante la conexión *batchClassifier*.

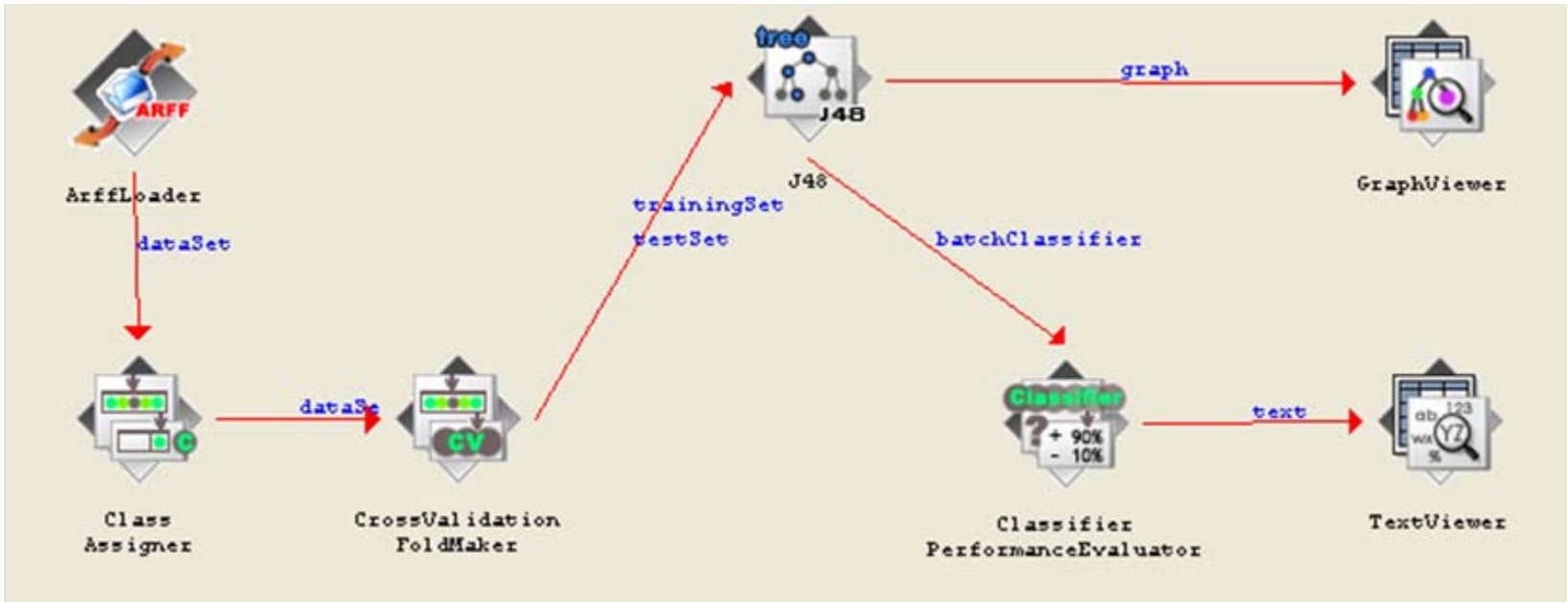
Introducción a WEKA

Knowledge Flow: Ejemplo paso a paso (7)

- Para finalizar, incluimos un *GraphViewer* y un *TextViewer* conectados a *J48* y *Classifier PerformanceEvaluator* mediante conexiones *graph* y *text*, respectivamente.
- Para ejecutar todo el experiento, seleccionar *Start Loading* en *Arff Loader.*
- Los resultados pueden verse en cualquiera de los dos visores finales

Introducción a WEKA

Knowledge Flow: Grafo final del ejemplo



Introducción a WEKA

Knowledge Flow (componentes más útiles)

- Visualización
  - DataVisualizer: Visualiza datos en 2D
  - AtributteSummarizer: Histogramas, uno por atributo
  - ModelPerformanceChart: Curvas ROC
  - TextViewer: Visualiza datos o modelos en texto
  - GraphViewer: Visualiza modelos de árboles
- Evaluación
  - CrossValidationFoldMaker: Divide datasets en folds
  - TrainTestSplitMaker: Divide un dataset en train/test
  - ClassAsigner: Asigna un atributo como clase
  - ClassValuePicker: Elige un valor como clase positiva
  - ClassifierPerformanceEvaluator: Recolecta estadísticas para evaluación batch.
  - IncrementalClassifierEvaluator: Recolecta estadísticas para evaluación incremental.
  - ClustererPerformanceEvaluator: Recolecta estadísticas para clustering
  - PreddictionAppender: añade predicciones de un clasificador a un dataset

Preprocesamiento en WEKA

WEKA contiene métodos de preprocesamiento para tratar valores perdidos, transformar datos, discretizar y seleccionar características e instancias

Todos estos métodos están en el apartado *Filters*. La selección de características la trata aparte.

WEKA no contiene métodos de selección de instancias, pero pueden simularse.

Preprocesamiento en WEKA

Filters (algoritmos más conocidos)

- Utilidades

Add: Añade un nuevo atributo

AddCluster: añade un atributo nominal para representar clusters

AddNoise: Cambia un porcentaje de valores de un atributo

Remove: Borra atributos

RemoveType: Borra atributos de un tipo (nominal, real,…)

SwapValues: Intercambia dos valores en un atributo

ClassOrder: Desordena el orden de los valores de clase

Randomize: Desordena el orden de las instancias

StratifiedRemoveFolds: Devuelve 1 fold de un dataset

RemovePercentage: Borra un porcentaje del dataset

RemoveRange: Borra un rango de instancias

RemoveWithValues: Borra instancias con ciertos valores

Normalize: Escalado de atributos numéricos a un intervalo

Preprocesamiento en WEKA

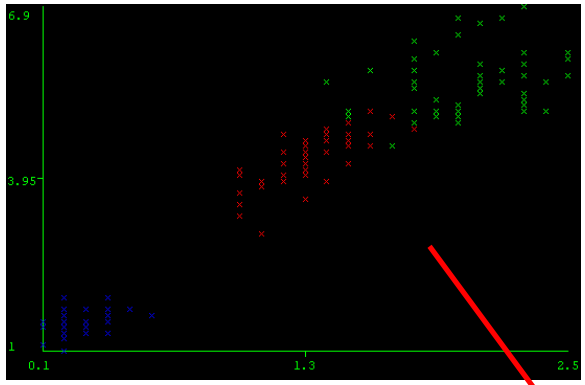
Filters (algoritmos más conocidos)

- Transformaciones
  - Normalize: Escalado de atributos numéricos a un intervalo
  - NominalToBinary: Transforma valores nominales a binarios
  - RandomProjection: Proyecta los datos en dimensión n a datos en dimensión m, siendo m < n
  - Standarize: Estandariza valores número a media 0 y desviación típica 1
- Tratamiento de valores perdidos
  - ReplaceMissingValues: Sustituye todos los valores perdidos para atributos nominales y numéricos con las modas y medias de los datos de entrenamiento

Preprocesamiento en WEKA

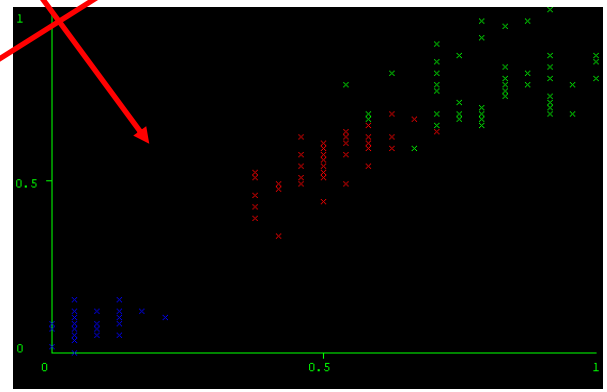
Ejemplo de Transformación

- Normalizado
  - *Filters*
  - *Unsupervised*
  - *Atribute*
  - *Normalize*



No.	sepallength Numeric	sepalwidth Numeric	petallength Numeric	petalwidth Numeric	class Nominal
1	5.1	3.5	1.4	0.2	Iris-se...
2	4.9	3.0	1.4	0.2	Iris-se...
3	4.7	3.2	1.3	0.2	Iris-se...
4	4.6	3.1	1.5	0.2	Iris-se...
				1.4	0.2 Iris-se...
6	5.4	3.9	1.7	0.4	Iris-se...
7	4.6	3.4	1.4	0.3	Iris-se...
8	5.0	3.4	1.5	0.2	Iris-se...
9	4.4	2.9	1.4	0.2	Iris-se...

Right click (or left+alt) for context menu



No.	sepallength Numeric	sepalwidth Numeric	petallength Numeric	petalwidth Numeric	class Nominal
1	0.222222...	0.624999...	0.067796...	0.041666...	Iris-se...
2	0.166666...	0.416666...	0.067796...	0.041666...	Iris-se...
3	0.111111...	0.5	0.050847...	0.041666...	Iris-se...
4	0.083333...	0.458333...	0.084745...	0.041666...	Iris-se...
5	0.194444...	0.666666...	0.067796...	0.041666...	Iris-se...
6	0.305555...	0.791666...	0.118644...	0.125000...	Iris-se...
7	0.083333...	0.583333...	0.067796...	0.083333...	Iris-se...
8	0.194444...	0.583333...	0.084745...	0.041666...	Iris-se...
9	0.027777...	0.374999...	0.067796...	0.041666...	Iris-se...

Reducción de Datos en WEKA

Discretización

- *Discretize*: Convierte atributos numéricos a nominales.
  - Especificar qué atributos, número de intervaloes, optimización de los mismos. Intervalos de igual anchura o frecuencia.

- *PKIDiscretize*: Discretiza con intervalos de igual frecuencia, y el número de intervalos es igual a la raiz cuadrada del número de valores.

Reducción de Datos en WEKA

Selección de Características

- La Selección de Características se realiza haciendo una búsqueda en el espacio de subconjuntos de características y evaluando cada uno de ellos.

- Se consigue combinando uno de los 4 evaluadores de subconjuntos con alguno de los 7 métodos de búsqueda implementados.

Reducción de Datos en WEKA

Selección de Características

- Evaluador de subconjuntos
  - CfsSubsetEval: Considera el valor predictivo individual de cada atributo
  - ClassifierSubsetEval: Usar un clasificador para evaluar
  - ConsistencySubsetEval: Mide la consistencia en términos de las clases
  - WrapperSubsetEval: Usa un clasificador + validación cruzada
- Métodos de búsqueda
  - BestFirst: Greedy Incremental con backtracking
  - ExhaustiveSearch: Fuerza bruta
  - GeneticSearch: Algoritmo genético de búsqueda
  - GreedyStepWise: Greedy incremental sin backtracking
  - RaceSearch: Metodología RaceSearch
  - RandomSearch: Búsqueda Aleatoria
  - RankSearch: Ordena los atributos y crea un ranking de subconjuntos prometedores.

Reducción de Datos en WEKA

Selección de Características

- Un método más rápido pero menos preciso consiste en evaluar los atributos individualmente y ordenarlos, descartando atributos que caen debajo de un determinado umbral.

- Se consigue seleccionando uno de los ocho evaluadores de atributos simple, usando después el método *Ranker* en la búsqueda.

Reducción de Datos en WEKA

Selección de Características

- Evaluadores de atributos simples

ChiSquaredAttributeEval: Calcula la estadística chi-cuadrado de cada atributo con respecto a la clase

GainRatioAttributeEval: Evaluación por tasa de ganancias

InfoGainAttributeEval: Evaluación por ganancia de información

OneRAttributeEval: Metodología OneR

PrincipalComponents: Análisis de principales componentes y transformación

ReliefFAttributeEval: Evaluados basado en instancias

SVMAttributeEval: Usar máquinas de soporte vectorial para calcular los atributos

SymmetricalUncertAttributeEval: Evalúa atributos basándose en incertidumbre simétrica

Reducción de Datos en WEKA

Selección de Instancias

- Los métodos clásicos de selección de instancias no están incluidos en WEKA.
- Solamente tiene incluido el **Muestreo Aleatorio**.
  - Se establece un porcentaje final de ejemplos
  - *Filter -> Unsupervised -> Instance -> Resample*
    - No mantiene la proporción de clases
  - *Filter -> Supervised -> Instance -> Resample*
    - Mantiene la proporción de clases

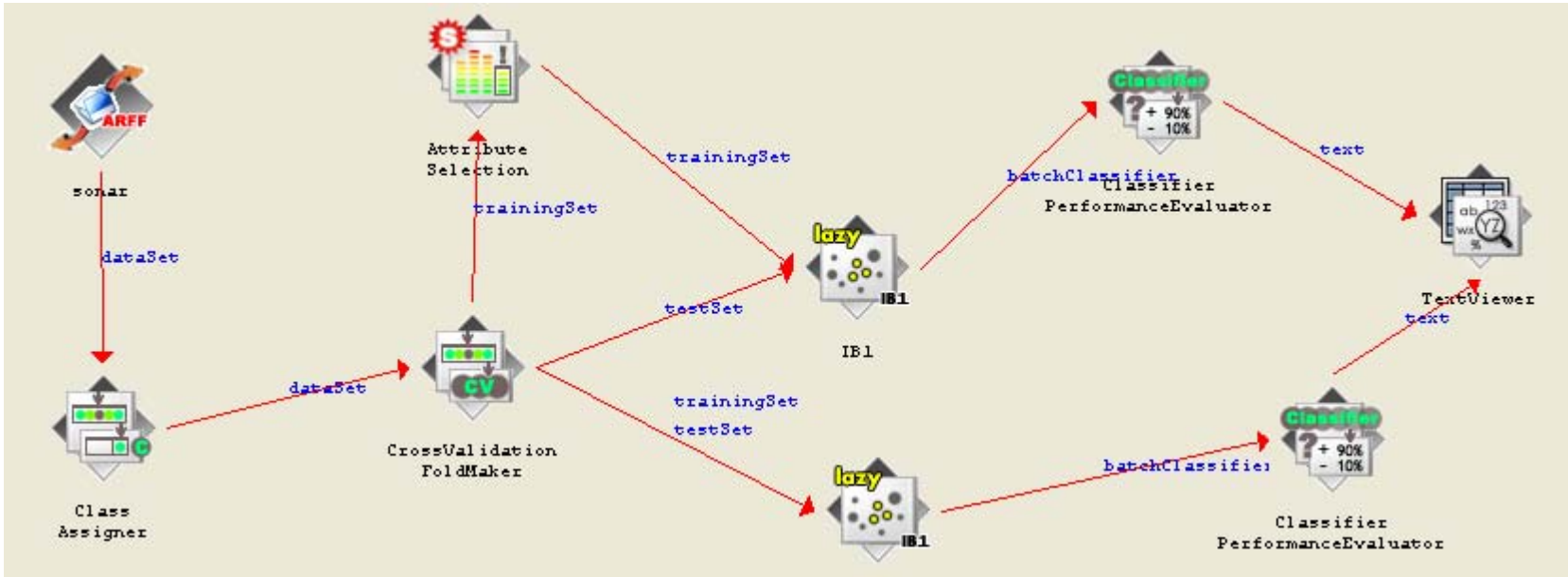
Reducción de Datos en WEKA

Selección de Instancias

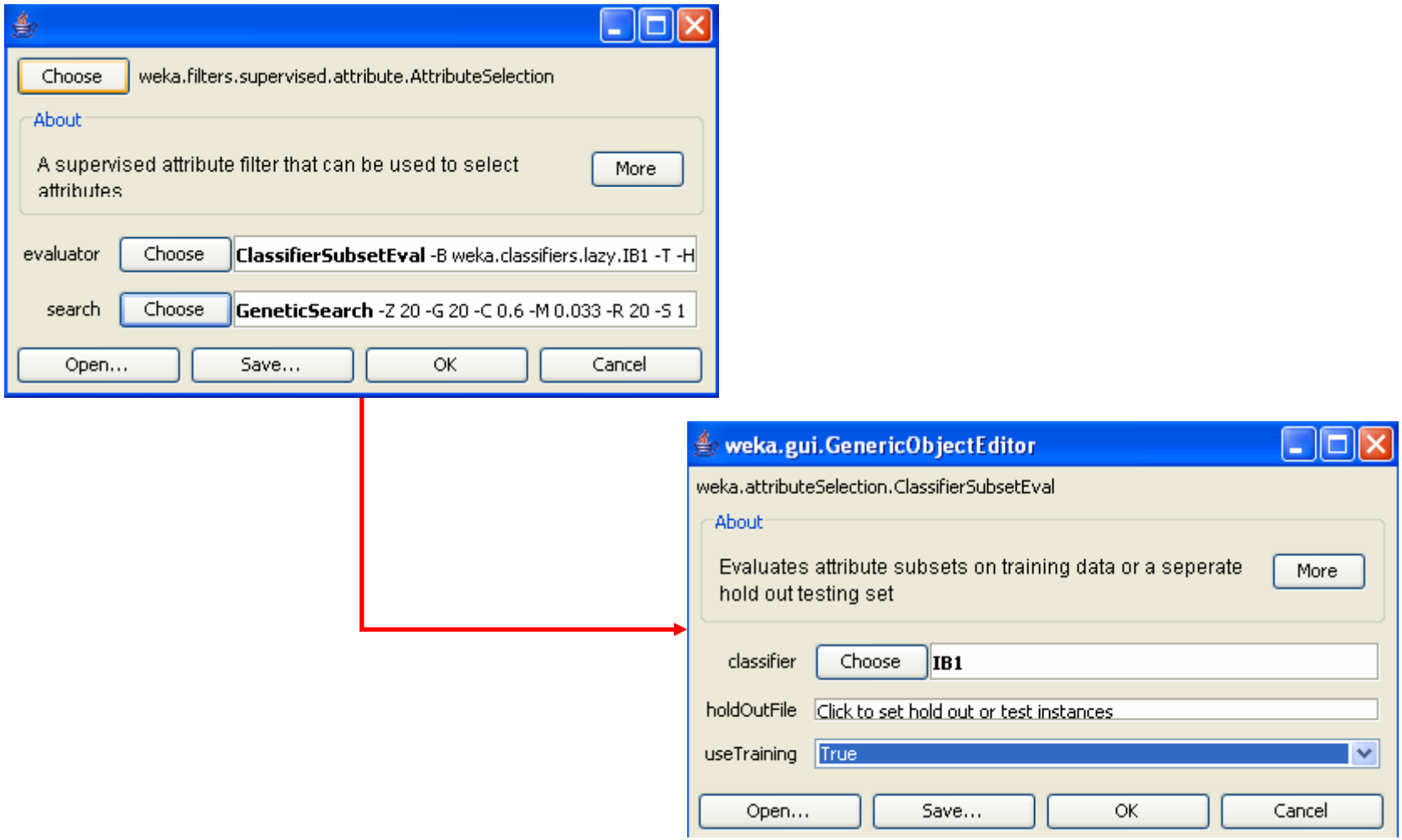
- Se pueden simular dos métodos clásicos de selección de instancias aplicando un clasificador y quitando o dejando los ejemplos clasificados incorrectamente.
- *Filters -> Unsupervised -> Instance -> RemoveMisclassified*
  - *CNN (Condensed Nearest Neighbour): Ib1* con *invert = false*
  - *ENN (Edited Nearest Neighbour): Ib3* con *invert = true*

Ejemplo 1

- Selección de características haciendo una búsqueda con un algoritmo genético del subconjunto de características y evaluando con Ib1 con el dataset *sonar*. Comparamos con Ib1 sin selección anterior.

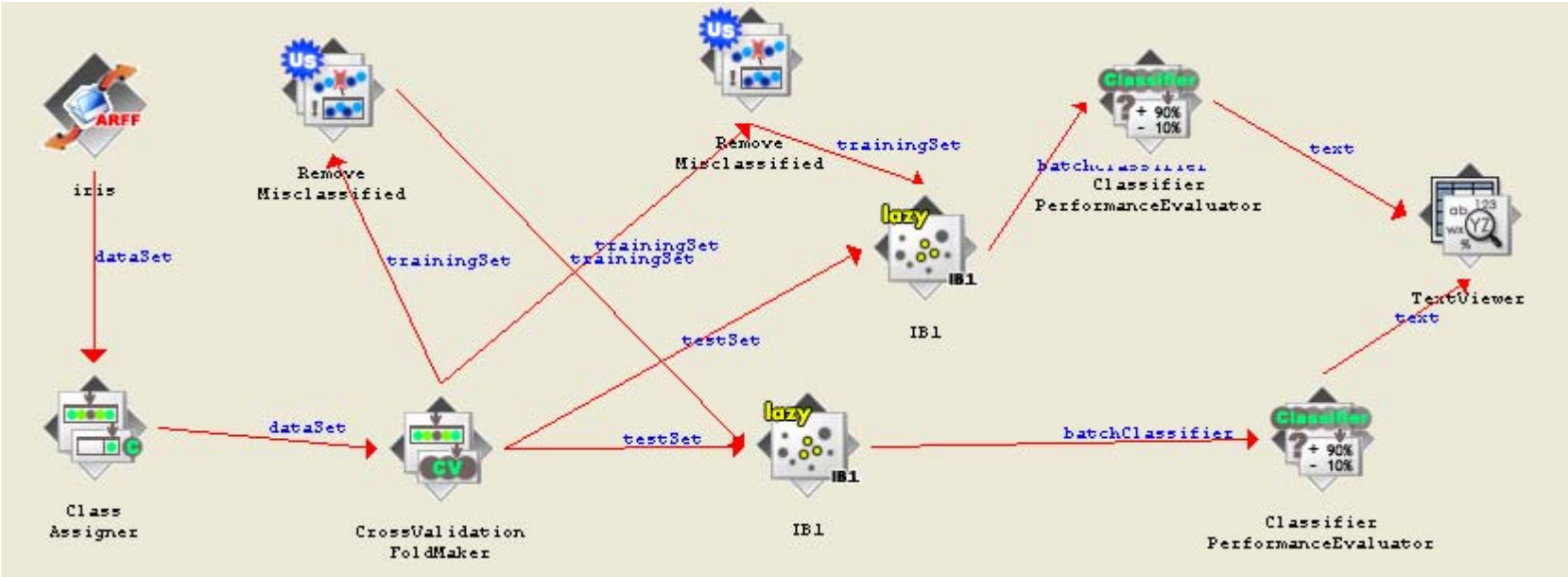


Ejemplo 1

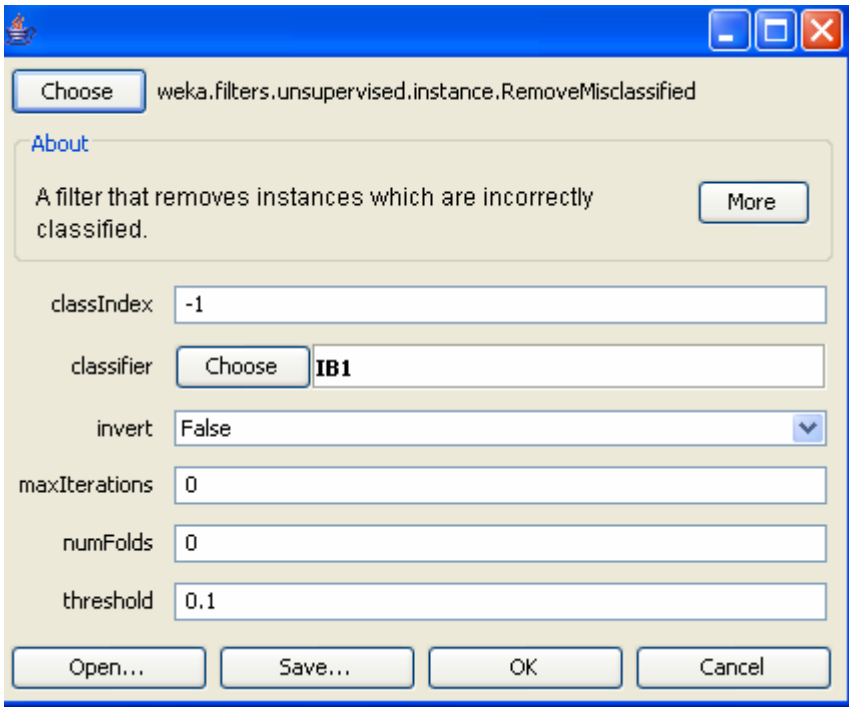

Choose
weka.filters.supervised.attribute.AttributeSelection
About
A supervised attribute filter that can be used to select attributes
More
evaluator
Choose
ClassifierSubsetEval -B weka.classifiers.lazy.IB1 -T -H
search
Choose
GeneticSearch -Z 20 -G 20 -C 0.6 -M 0.033 -R 20 -S 1
Open...
Save...
OK
Cancel
weka.gui.GenericObjectEditor
weka.attributeSelection.ClassifierSubsetEval
About
Evaluates attribute subsets on training data or a seperate hold out testing set
More
classifier
Choose
IB1
holdOutFile
Click to set hold out or test instances
useTraining
True
Open...
Save...
OK
Cancel

Ejemplo 2

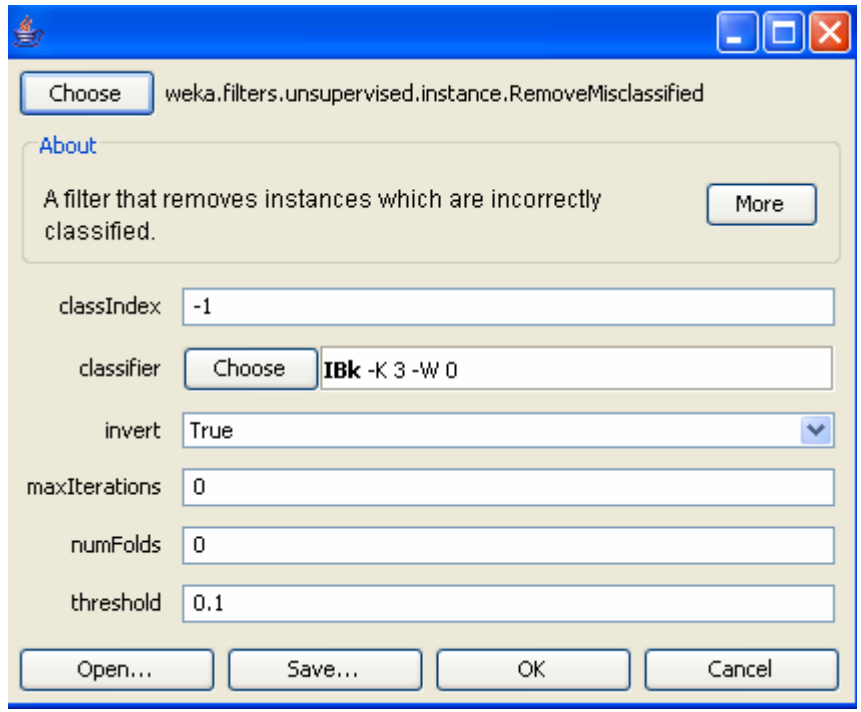
- Selección de instancias con CNN y ENN sobre *Iris*.



Ejemplo 2



CNN



ENN

Ejercicios

- 1.- Hay métodos que solo funcionan con datos de tipo nominal. Éste es el caso de ID3. Para trabajar con este algoritmo, hay que discretizar todos los atributos reales. En Explorer y utilizando el dataset *glass*, consigue ejecutar ID3 con una discretización previa. Los datasets “extra” los puedes conseguir desde http://sci2s.ugr.es/keel/UCI.zip

Ejercicios

- 2.- Combinar diferentes estrategias de Selección de Características e Instancias con Knowledge Flow utilizando como clasificador *J48* y cualquier dataset no estándar en WEKA.

http://sci2s.ugr.es

Gracias !!!