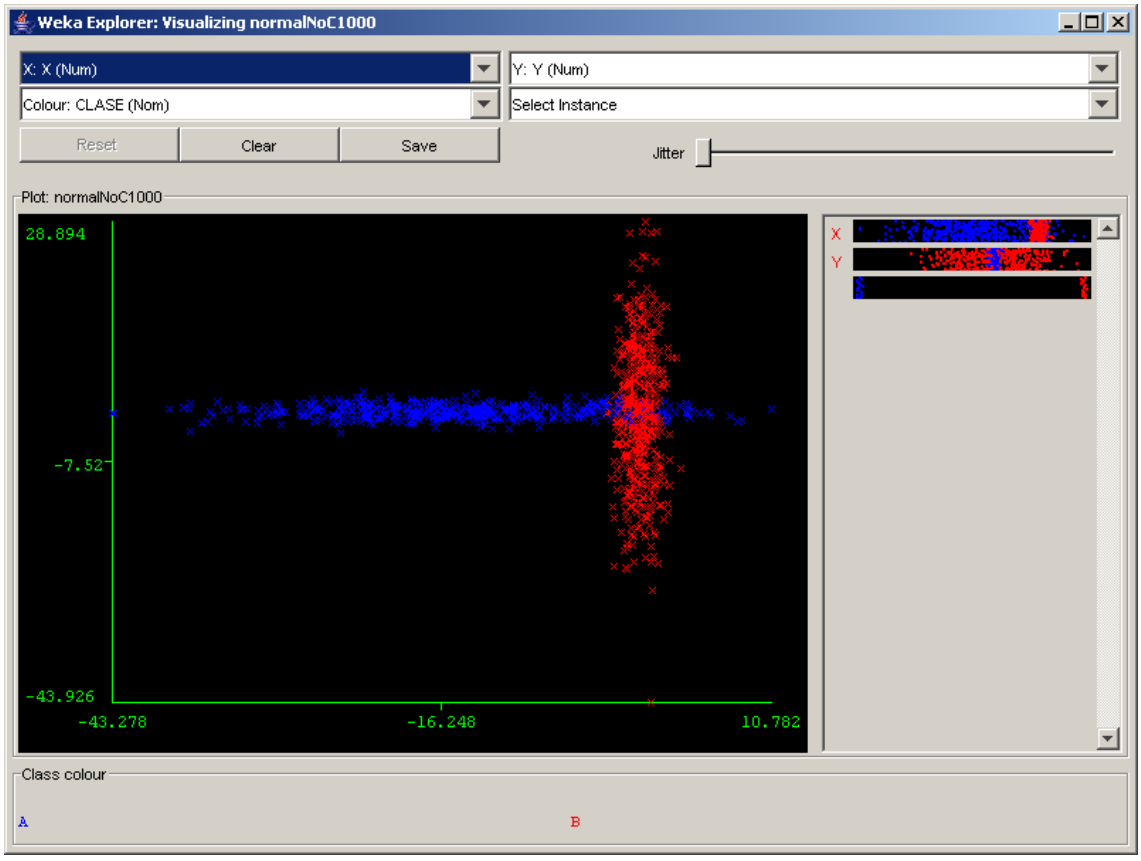
Universidad Carlos III de Madrid

Análisis de Datos

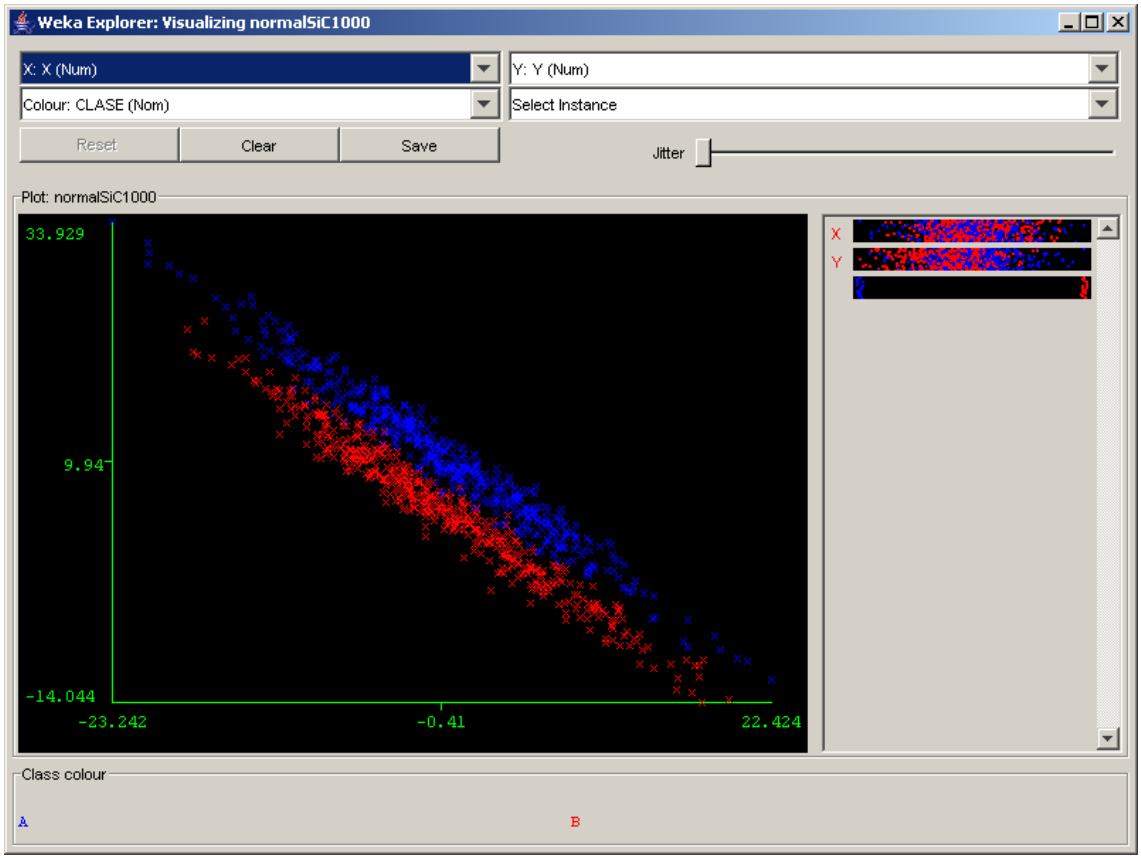
Jesús García Herrero

Práctica 3 WEKA. Comparación de clasificadores con WEKA

En este ejercicio se propone llevar a cabo una comparación de distintos clasificadores sobre varios conjuntos de datos, mediante la técnica estadística de tipo test pareado (*t-test*). Para ello se utilizará la herramienta *experimenter* de WEKA, sobre datos generados a partir de variables normales bidimensionales. Los dos tipos de situaciones estudiadas se presentan en las figuras siguientes, que se corresponden con variables independientes en ambas coordenadas (NoC) o con covarianza no nula (SiC).



datosNoC1000.arff



datosSiC1000.arff

Los ficheros *datosSiC1000.arff, datosNoC1000.arff* contienen 1000 instancias, y los ficheros *datosSiC200.arff* y *datosNoC200.arff* 200. Los calificadores "SiC", "NoC" significan, respectivamente, sin correlación y con correlación. Se van a evaluar los clasificadores de tipo "naive bayes", "C4.5" y clasificación mediante regresión lineal. Se ejecutarán 10 validaciones cruzadas (con 10 partes cada una), y se promediarán los resultados, determinándose qué técnica es mejor para un cierto nivel de confianza, con un test de tipo t-student pareado. Para ello:

- Entrar en WEKA, y seleccionar "Experimenter".
- Dentro de SETUP la opción "Advance" dentro de Experiment Configuration Mode.
- Crear un nuevo experimento (Botón "NEW").
- Seleccionar el destino del experimento.
    - Seleccionar InstantesResultListener.
    - Seleccionar el fichero de salida (output file) "experimento01".
- Seleccionar el generador de resultados.
    - CrossValidationResultProducer..
- Seleccionar el número de ejecuciones (RUNS)(1 to10).

• En el panel de Generador de Propiedades seleccionar “ENABLE” -> Select Property -> SplitEvaluator -> Classifier y botón “SELECT”.

• Seleccionar los tres clasificadores a analizar. (Para clasificación mediante regresión, seleccionar el metaclasificador “ClassificacionViaRegression”, utilizando la función “LinearRegression” como función de predicción).

• En el panel de *Datasets* seleccionar los cuatro ficheros a analizar.

• Seleccionar pestaña de “RUN".

• Seleccionar botón “START".

• Seleccionar botón “SAVE” para guardar la definición del experimento. Guardarlo como “experimento01.exp”.

• Cuando finalice el experimento, seleccionar la pestaña “ANALYZE"

• Seleccionar botón “EXPERIMENT"

• Seleccionar botón “PERFORM TEST". Variar el clasificador usado como base y el intervalo de confianza.

• Esta misma técnica puede utilizarse para analizar diferentes parámetros de un clasificador para determinar el conjunto más adecuado. Por ejemplo, 3 clasificadores de tipo C4.5, variando el intervalo de confianza para podar con los valores {0.25, 0.1, 0.01}.