Análisis de datos con Weka

Sistemas Inteligentes 2022-2023

El objetivo de esta práctica es usar Weka para resolver un problema de análisis de datos. Cada alumno trabajará con un conjunto de datos diferente. Para esto se proporciona una carpeta con diferentes conjuntos de datos, los cuales tienen como nombre de fichero el número de grupo de prácticas en el que se inscribieron a través de la plataforma de evaluación virtual. *Si por alguna circunstancia no puede trabajar con el fichero asignado, presenta alguna dificultad inesperada, o simplemente considera que el conjunto de datos no es lo suficientemente complejo como para probar las cosas que le gustaría, póngase en contacto con el profesor.*

En esta práctica se espera que el alumno investigue nuevas capacidades de Weka no vistas en clase, de modo que pueda desenvolverse en un entorno cambiante en el que por muchas actividades que desarrollemos en prácticas el día de mañana se va a encontrar con muchas necesidades adicionales a las que deberá responder. Por tanto, se tendrá en cuenta en la evaluación de la práctica el uso de esas capacidades yendo más allá de las opciones específicas del software que explícitamente se vieron en clase, explicando en el informe los nuevos hallazgos que han emergido de la exploración de las herramientas empleadas y posibles recursos adicionales.

Emplee el algoritmo J48 junto a otros clasificadores presentes en Weka y escriba un informe sobre el fichero y sobre el proceso de aprendizaje realizado. Este informe incluirá en la medida de lo posible, información sobre el conjunto de datos utilizado, visualización/exploración (si se considera que hay algo interesante que mostrar), transformaciones aplicadas sobre variables (normalización, discretización, imputación de valores perdidos, etc.), documentación sobre cada una de las pruebas realizadas y las puntuaciones obtenidas, metodología de validación usada en los diferentes experimentos (train, train/test, validación cruzada, etc.).

Recordemos que un proyecto de análisis de datos es un proceso iterativo en el que vamos realizando experimentos, con el objetivo de obtener el mejor modelo predictivo posible (profundizaremos más en la próxima clase). Al final se deberá reflexionar sobre los diferentes resultados obtenidos en cada prueba y por qué unos son mejores que otros. A continuación se enumeran algunas de las tareas típicas que se realizan en un proyecto de análisis de datos con el objetivo de obtener resultados cada vez mejores:

- Exploración/visualización (en este caso el objetivo es entender el problema y sus datos)
- Imputación de valores perdidos (esto será necesario para algunos conjuntos de datos que contienen valores faltantes)
- Normalización, estandarización, discretización, codificación one-hot, etc.
- Creación de nuevos atributos como se vio en la práctica

- Selección de atributos
- Selección de los mejores parámetros de configuración de un algoritmo
- Selección del mejor algoritmo

Como ayuda para empezar a explorar, se proporciona el equivalente en Weka de los algoritmos vistos (o que se verán) en teoría, sin embargo, no tiene por qué limitarse a estos ni tampoco usarlos todos.:

- ID3/CART – J48 (classifiers/trees)
- Aprendizaje de reglas por cobertura - No está presente en Weka aunque proporciona algoritmos similares (classifiers/rules)
- Clasificador Naive Bayes – NaibeBayesMultinomial (classifiers/bayes) – Nótese que este algoritmo no es aplicable sobre cualquier tipo de datos.
- KNN – IBK (classifiers/lazy)
- Kmedias – SimpleKMeans (pestaña clusters/clusterers). Cuidado, esta técnica, al contrario que las anteriores es de aprendizaje NO supervisado.
- Perceptrón Multicapa (tema de redes neuronales) – MultilayerPerceptron (classifiers/functions)

Además de los algoritmos y sus parámetros, puede ser interesante probar transformaciones sobre variables ya que según le proporcionemos los datos al algoritmo, es posible que que este sea capaz de explotar mejor la información (esto se haría en la pestaña preprocess de Weka). Algunos ejemplos son la 'binarización' de variables 'categóricas' (también conocido como OneHotEncoding) o la discretización de variables continuas (convertir variables continuas en categorías).

Se recomienda comenzar revisando la práctica 5. Adicionalmente, existe gran cantidad de documentación en Internet sobre Weka que nos ayudará a entender cada uno de los procesos que podemos realizar.
Cada fichero de datos puede ser abierto con un editor de texto plano (como el blog de notas, Notepad++, etc.) para poder ver su contenido completo. La mayoría de los ficheros contienen, además de los datos en formato Weka, documentación sobre el origen y propósito del conjunto de datos.

Variable respuesta: Algunos conjuntos de datos no especifican la variable respuesta (clase o variable a predecir). En este, caso podemos elegir alguna que sea predecible y/o tenga sentido predecir, si tiene dudas consulte al profesor. Adicionalmente, los alumnos cuyo conjunto de datos ya tiene una variable respuesta seleccionada, si lo desean, pueden probar a predecir otras variables. Weka por defecto usará la última variable del conjunto de datos (sea o no sea la variable respuesta) por lo que es recomendable leer la descripción del conjunto de datos para saber cuál es el objetivo. En algunos casos (no siempre) encontraremos información sobre la variable respuesta junto a las palabras clave CLASSTYPE y CLASSINDEX. En otros casos (no

siempre), la variable respuesta se proporciona con el nombre *class*. Un ejemplo de conjunto de datos que no tiene definida la variable a predecir es:
% CLASSTYPE: nominal
% CLASSINDEX: none specific

Para cualquier duda sobre el conjunto de datos recibido o sobre Weka póngase en contacto con el profesor. Adicionalmente se podrá solicitar de forma justificada un cambio de conjunto de datos (por ejemplo, *tengo un conjunto de datos con pocas variables y me gustaría profundizar en la exploración de técnicas de selección de atributos*).

Entregable: Un fichero en formato ***pdf*** con un informe sobre el trabajo realizado y los resultados obtenidos.

- Se valorará la completitud del estudio, el uso de nuevas técnicas y la interpretación de los resultados obtenidos.
- Si las capturas de pantalla que incluya en el entregable contienen texto, asegúrese de que sea legible.
- Incluya los resultados obtenidos en los diferentes experimentos realizados junto con la configuración utilizada (algoritmo seleccionado y parámetros que ha modificado). Se recomienda proporcionar un resumen en forma de tabla.

MUY IMPORTANTE:

Cualquier plagio que se detecte en **cualquier parte del trabajo** significará automáticamente la **calificación de cero** en la asignatura para **todos los alumnos involucrados** (la responsabilidad de la copia de una práctica es compartida entre los alumnos). Por tanto, a estos alumnos **no se les conserva**, ni para la actual ni para futuras convocatorias, **ninguna nota** que hubiesen obtenido hasta el momento. Todo ello **sin perjuicio de las** correspondientes **medidas disciplinarias** que se pudieran llevar a cabo.