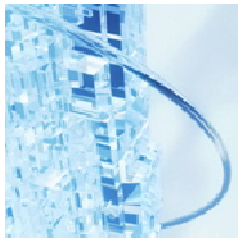

Prácticas de Minería de Datos

Grado en Ingeniería Informática

Curso 2013-14

PRÁCTICA FINAL

Proceso Completo de Extracción de Conocimiento en BD

Proyecto de KDD
(Knowledge Discovery in Databases)

OBJETIVOS

- Estudiar e implementar un proyecto completo de extracción de conocimiento en BD (KDD), cubriendo todas sus fases
- Familiarizar al alumno con la realización y presentación de proyectos de Minería de Datos

1. Planteamiento del Proyecto

Se propone al alumno la resolución de un problema de clasificación mediante la realización (de forma individual) de un proyecto completo de KDD en la que se cubran todas las etapas del proceso de extracción de conocimiento (Integración y recopilación, Selección de datos, limpieza y transformación, Selección de la tarea y técnicas de MD a aplicar, Evaluación, interpretación y presentación de los resultados obtenidos).

Para la realización del proyecto, se proponen 5 conjuntos de datos de problemas reales obtenidos del repositorio UCI, de los cuales cada estudiante deberá seleccionar uno de los datasets propuestos.

Para cada dataset se han confeccionado 2 ficheros: uno de entrenamiento (80% de las instancias) y otro de test (20% de las instancias), realizando una selección estratificada de los datos, a fin de distribuir de manera uniforme las clases entre ambos ficheros. El fichero de test se deberá utilizar en la fase final del proyecto y sólo podrá ser usado para validar los resultados obtenidos durante la fase de entrenamiento.

2. Desarrollo del Proyecto

El desarrollo del proyecto implica la realización de todas las fases del proceso de KDD. La evaluación del proyecto no dependerá del nivel de bondad de los resultados medidos, sino del nivel de desarrollo del proceso de KDD.

Los pasos que se deberían seguir son:

1. **Comprensión del problema**: Se deberá conocer el problema que se plantea, los atributos que proporciona el dataset y lo que significa/indica cada uno de ellos y, finalmente, el objetivo de la clasificación.
2. **Selección y preparación de los datos**: Se construirán los ficheros con el formato adecuado a la herramienta que se vaya a utilizar (en nuestro caso WEKA).
3. **Identificación de la problemática intrínseca en los datos disponibles**: existencia de valores ausentes, outliers, desbalanceo, etc.
4. **Limpieza y preprocesado**: En esta fase, según los problemas detectados en el punto 2, se deberá actuar en consecuencia.
5. **Transformación**: Dependiendo de la naturaleza del problema, se estudiará la oportunidad de generación de nuevas características, normalización de valores, etc.

6. **Minería de datos**: Se elegirán los métodos/algoritmos que se consideren más oportunos dada las características del problema. Al menos se deberán probar cinco modelos distintos. Se considerará la oportunidad de probar los modelos con distintos tratamientos de preprocesado y/o transformación.
7. **Evaluación e interpretación de resultados**: Se evaluarán los resultados con las medidas propuestas y se obtendrán las conclusiones oportunas.

3. Normativa de entrega del proyecto

- Se deberá entregar una memoria descriptiva de los pasos seguidos en el proceso de *KDD* para la obtención de los modelos de clasificación (el número mínimo de modelos a desarrollar es cinco) generados para el *dataset* objeto de estudio.

- La memoria del proyecto se debe realizar utilizando la plantilla LNCS. Se recomienda que tenga un máximo de 12 páginas (incluidas tablas y figuras). Podéis descargar la plantilla desde la página de la asignatura.

- La memoria deberá seguir un esquema similar al que os proponemos:

    1. **Título del trabajo**. Debéis proponer un título donde queden reflejadas las características más destacadas del trabajo. El título deberá indicar el problema abordado, que viene dado por el conjunto de datos estudiado.
    2. **Resumen / Abstract** (máximo 6 líneas en español e inglés).
    3. **Descripción del problema y del conjunto de datos**. Presentará tanto la problemática del problema al que representa el conjunto de datos, como la problemática intrínseca de los propios datos (desbalanceo, valores ausentes, outliers, multiclase, ...)
    4. **Preparación de los datos.** Selección de filas y características, limpieza, balanceo de clases, etc.
    5. **Descripción de los atributos seleccionados.** En esta sección se puede hacer un estudio, mediante técnicas de visualización, de la distribución de los valores de algunos de los atributos y su posible influencia en los algoritmos que se han aplicado.
    6. **Descripción de los algoritmos de clasificación elegidos para hacer el estudio.** Se pueden describir los parámetros elegidos, el método de evaluación, etc.
    7. **Evaluación de los resultados obtenidos con el conjunto de entrenamiento.** Se evaluarán las siguientes medidas: porcentaje de acierto, medida-F y AUC. Éstas se deben obtener mediante 5-validación cruzada sobre el conjunto de entrenamiento para cada uno de los modelos desarrollados.
    8. **Evaluación de los resultados obtenidos con el conjunto de test.** Se evaluarán las mismas medidas obtenidas sobre el conjunto de test para cada uno de los modelos desarrollados.
    9. Conclusiones y valoración del proyecto **(OBLIGATORIO)**

- Además de la memoria, se deberá preparar una presentación, de un máximo de 10 minutos, para exponer el proceso desarrollado y los logros obtenidos. La presentación se expondrá en clase el día de la defensa ante el resto de compañeros.

¿Cómo entregar la práctica?

- Para subir el proyecto se deberá crear un archivo comprimido que contenga la memoria y la presentación.
- El fichero comprimido se subirá a través de la plataforma web.
- Fecha máxima de entrega: **martes, 4 de marzo de 2014, hasta las 10:00**
- Fecha de la defensa y discusión de resultados: **a consensuar con los alumnos**