UNIVERSIDAD DE GRANADA
E.T.S. de Ingenierías Informática y de Telecomunicación

Departamento de Ciencias de la Computación e Inteligencia Artificial

Inteligencia de Negocio

Guión de Prácticas

Práctica 1: Análisis Predictivo Empresarial Mediante Clasificación

Curso 2018-2019

Cuarto Curso del Grado en Ingeniería Informática

Práctica 1

Análisis Predictivo Empresarial Mediante Clasificación

1. Objetivos y Evaluación

En esta primera práctica de la asignatura Inteligencia de Negocio veremos el uso de algoritmos de aprendizaje supervisado de clasificación como herramienta para realizar análisis predictivo en la empresa. En ella el alumno adquirirá capacidades para abordar problemas reales donde la minería de datos puede aportar valor en forma de conocimiento para ayudar en la toma de decisiones para gestión empresarial. Concretamente, se trabajará con un conjunto de datos real sobre el que se emplearán diferentes algoritmos de clasificación (para su comparación) y a la luz del conocimiento descubierto se podrán concluir estrategias para resolver el problema. Para ello, se deberán crear informes de resultados y análisis lo suficientemente profundos para resultar de utilidad.

La práctica se calificará hasta un **máximo de 2 puntos**. Se valorará el acierto en los recursos de análisis gráficos empleados, la complejidad de los experimentos realizados, la interpretación del conocimiento extraído, la organización y redacción del informe, etc.

2. Descripción del Problema: Predicción de la Popularidad de Noticias Online

La alta proliferación de espacios con noticias en la Web hace cada vez más útil la predicción automática de la popularidad de estas noticias. En esta práctica se propone un sistema inteligente de apoyo a la decisión (SIAD) que analiza los artículos antes de su publicación. Se usa un amplio conjunto de características extraídas (por ejemplo, palabras clave, contenido de medios digitales, popularidad anterior de las noticias a las que se hace referencia en el artículo, etc.). El SIAD predice si un artículo se volverá popular. En un caso real, esta herramienta puede combinarse con un algoritmo de optimización para mejorar las características del artículo y hacerlo así más popular. El conjunto de datos se basa en 39.644 noticias extraídas en 2015 de

la web `http://mashable.com`. Además de las características extraídas sobre estas noticias, se conoce también el número de veces que la noticia ha sido compartida. A partir de este dato, se puede valorar si la noticia es popular o no. En nuestro caso, fijaremos como umbral 3000 veces, de modo que si la noticia se comparte en un número superior a ese umbral, la noticia es considerada popular.

El conjunto de datos está disponible en la web de la asignatura `http://sci2s.ugr.es/graduateCourses/in` (el cual se ha obtenido mediante algunas transformaciones sobre el original disponible en `http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity`). Se recomienda la lectura del siguiente artículo para conocer mejor el problema: "*K. Fernandes, P. Vinagre and P. Cortez. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal*".

3. Tareas a Realizar

La práctica consiste principalmente en que el alumno estudie el comportamiento de distintos algoritmos de clasificación mediante el diseño experimental apropiado y el análisis comparado de resultados. Además, también deberá extraer conclusiones a partir del conocimiento aprendido mediante estos algoritmos para comprender las relaciones entre las variables (también llamadas *características* o *predictores*) que favorecen una determinada clase. El trabajo se realizará sobre la plataforma KNIME (`http://www.knime.org`), pudiéndose emplear nodos adicionales de extensiones tales como Weka, JFreeChart o JavaScript Views.

Concretamente, se deberán resolver adecuadamente las siguientes tareas:

1. Se considerarán al menos seis algoritmos de clasificación distintos. Se valorará la selección justificada de estos algoritmos en función de las características del conjunto de datos así como la elección de variedad de tipos de representación (árboles, reglas, redes neuronales, etc.).

2. Toda la experimentación se realizará con validación cruzada de 5 particiones. Para sustentar el análisis comparativo se emplearán tablas de errores, matrices de confusión y curvas ROC. Además de la precisión, se añadirán otras medidas de rendimiento como TPR, TNR, AUC, Valor-F_1, G-mean y de complejidad del modelo (número de hojas, número de reglas, número de nodos, etc.).

3. Todos los análisis de resultados serán comparativos, de forma que se estudien los pros y contras de cada representación y/o de cada algoritmo. La documentación deberá incluir al menos una tabla resumen que incluya los resultados medios de todos los algoritmos analizados. El análisis no podrá reducirse a una simple lectura de los resultados obtenidos. El alumno deberá formular y argumentar hipótesis sobre las razones de cada resultado. En este problema, ¿por qué el algoritmo X funciona mejor que el Y? ¿Por qué la representación X presenta ciertas ventajas respecto a la Y?

4. Se probarán configuraciones alternativas de los parámetros de los algoritmos empleados justificando los resultados obtenidos. Por ejemplo, ¿puedo evitar o paliar el sobreaprendizaje ajustando los parámetros? ¿Puedo obtener modelos más fácilmente interpretables sin sacrificar excesiva precisión? Para realizar este análisis, se incluirán tablas comparativas con los resultados del algoritmo con parámetros o configuración por defecto y con las distintas variaciones estudiadas. Si el análisis es suficientemente completo, no es necesario estudiar todos los algoritmos analizados, se pueden escoger solo algunos de ellos.

5. Se deberán analizar los datos con diferentes gráficas para comprender su naturaleza e influencia en el proceso de clasificación.

6. A la luz de este análisis, se deberá estudiar un procesado básico de los datos que mejore la predicción (por ejemplo, eliminar alguna característica por razón justificada, agrupar los valores posibles de una característica, eliminar ciertas instancias del conjunto de entrenamiento que se consideren erróneas, convertir una característica categórica en varias binarias, imputar valores perdidos, equilibrar el balanceo de clases...). Deberán justificarse las acciones tomadas y analizar porqué determinado procesado funciona mejor en un determinado tipo de algoritmo. Si no se consigue mejorar la predicción, se podrá al menos describir los procesados que se han probado y los resultados obtenidos. De nuevo, se requiere una tabla resumen que muestre los resultados antes y después de los diferentes procesados de datos.

7. Basado en todo lo anterior, se deberán extraer conclusiones sobre los factores que determinan que una noticia sea popular. Para llegar a estas conclusiones, se pueden analizar los modelos legibles generados (por ejemplo, árboles de decisión, conjuntos de reglas o regresiones lineales), analizar la importancia de cada característica en el proceso de clasificación y visualizar los resultados de predicción de los modelos sobre diferentes casos de entrada (*What-If Analysis*).

4. Esquema de la Documentación

La documentación entregada deberá ajustarse al siguiente esquema (debe respetarse la numeración y nombre de las secciones):

1. **Introducción**: se hablará sobre el problema abordado y todas las consideraciones generales que se deseen indicar.

2. **Resultados obtenidos**: incluirá un apartado 2.x por cada algoritmo estudiado. En cada apartado se añadirán capturas de pantalla de KNIME que expliquen el flujo de trabajo empleado y una tabla con los resultados obtenidos por el algoritmo como se describe en la tarea 2.

3. **Análisis de resultados**: incluirá la tabla resumen de todos los algoritmos analizados así como su interpretación y análisis mencionados en la tarea 3. Se podrán añadir gráficas y visualizaciones que apoyen el análisis.

4. **Configuración de algoritmos**: se incluirá un apartado para cada algoritmo cuya configuración y parámetros hayan sido estudiados. En cada apartado, se incluirá una tabla con los resultados y se realizará el correspondiente análisis como se describe en la tarea 4.

5. **Procesado de datos**: se describirá el procesado realizado, la tabla de resultados y su análisis como se describe en la tarea 6. Se incluirán capturas de pantalla de KNIME con los flujos de trabajo usados para los distintos procesados.

6. **Interpretación de resultados**: como se describe en la tarea 7. Se incluirán las representaciones de modelos y visualizaciones de casos necesarias para sustentar la interpretación de resultados.

7. **Contenido adicional**: cualquier tarea adicional a las descritas en este guion puede presentarse en esta sección.

8. **Bibliografía**: referencias y material consultado para la realización de la práctica.

No se aceptarán otras secciones distintas de estas. Además, la primera página de la documentación incluirá una portada con el nombre completo del alumno, grupo de prácticas y dirección email. También se incluirá una segunda página con el índice del documento donde las diferentes secciones y páginas estarán enlazadas en el `pdf`.

5. Entrega

La fecha límite de entrega será el domingo **4 de noviembre** de 2018 hasta las **23:59**. La entrega se realizará a través de la web de la asignatura en `https://decsai.ugr.es`. En un único fichero `zip` se incluirá el árbol de directorios completo que contiene el proyecto de KNIME, la documentación de la práctica realizada en `pdf` y cualquier otro archivo que el alumno considere relevante. Para evitar un excesivo tamaño de archivo, el proyecto KNIME se puede entregar sin ejecutar (se puede hacer una copia del proyecto y resetear el primer nodo de lectura de datos) para que no se almacenen todos los resultados intermedios y por tanto se reduzca drásticamente su tamaño. No obstante, si, por alguna razón justificada, el archivo `zip` fuera demasiado grande para subirse en la plataforma, se podrá incluir en la documentación (visible en la primera página de portada) un enlace a `http://consigna.ugr.es` con el material adicional. En tal caso, la permanencia en consigna deberá ser de 1 mes y deberá comunicarse por email al profesor de prácticas. En ningún caso se aceptan otros tipos de enlaces como Dropbox, Google Drive o similares.

El nombre del archivo `zip` será el siguiente (sin espacios): `P1-apellido1-apellido2-nombre.zip`. La documentación tendrá el mismo nombre pero con extensión `pdf`. Es decir, la alumna “María

Teresa del Castillo Gómez" subirá el archivo `P1-delCastillo-Gómez-MaríaTeresa.zip` que contendrá, entre otros, el archivo `P1-delCastillo-Gómez-MaríaTeresa.pdf`.