PRÁCTICA DE INGENIERÍA DEL CONOCIMIENTO

CONSTRUCCIÓN DE UN SISTEMA IDENTIFICADOR DE SPAM MEDIANTE ÁRBOLES DE DECISIÓN

CONSTRUCCIÓN DE UN SISTEMA IDENTIFICADOR DE SPAM.

1. Descripción del problema

Ejemplos de spam son mensajes de e-mail no deseados como anuncios de sitios/productos web, formas de ganar dinero rápidamente, pornografía, e-mails comerciales no solicitados...
Los datos con los que entrenaremos el sistema se han recogido de miles de e-mails recibidos por empleados de diferentes empresas de Estados Unidos. Estos datos se pueden dividir en e-mails catalogados como spam y aquel conjunto de e-mails normales catalogados como no spam, provenientes en su mayoría de ficheros de trabajo y mensajes personales.
La técnica que se usa es identificar la aparición de palabras clave y secuencia de caracteres. La hipótesis de la que se parte es que en todos estos mensajes hay palabras tipo que suelen repetirse con cierta frecuencia, al igual que se repite la aparición de secuencias de caracteres características, como por ejemplo un gran número de mayúsculas o de consonantes seguidas (al estilo del spam que suele llagar a las cuentas de hotmail)
En el otro sentido hay palabras que claramente hacen que el mensaje no sea identificado como spam. Así "George" y el código de área 650 son indicadores de mensajes no-spam; esto puede ser útil por ejemplo cuando se trata de construir un filtro anti-spam personalizado.
En esta práctica obviaremos la tarea de recopilar mensajes y contar dicha secuencia de caracteres, y contaremos ya con estos datos. Nuestro objetivo es, haciendo uso de estos datos (aparición de palabras clave, secuencias de caracteres) identificar patrones de regularidad para poder detectar aquellos mensajes spam que recibamos después de entrenar al sistema.
Nuestra base de datos se compone de 400 ejemplos (160 Spam = 40%), mientras que el número de atributos es de 58 (57 continuos, 1 indicando la clase)

2. Descripción de los atributos:

- La última columna del fichero 'spambase.data' denota si el e-mail ha sido (1) o no (0) considerado como spam. La mayoría del resto de atributos indica si una palabra o carácter particular aparece o no en el mensaje. Los atributos (55-57) miden la longitud de secuencias de letras mayúsculas consecutivas. Todos ellos aparecen definidos en el fichero spambase.names.

  Las definiciones de los atributos son las siguientes:

  - 48 atributos reales continuos [0,100] de tipo word_freq_WORD= porcentaje de palabras en el e-mail que contienen la palabra WORD (entendida como palabra genérica), calculados de la siguiente forma: 100 * (número de veces que la palabra WORD aparece en el e-mail / número total de palabras en el e-mail. La palabra "WORD" en este caso es cualquier cadena de caracteres alfanuméricos rodeados por caracteres no-alfanuméricos o fin-de-cadena. Por ejemplo, word_freq_make indica la frecuencia de aparición de la palabra make.

- 6 atributos reales continuos [0,100] llamada char_freq_CHAR= porcentaje de caracteres en el e-mail que contienen CHAR, calculados de la siguiente forma: 100 * número de veces que la secuencia CHAR aparece / número total de caracteres en el e-mail.
- 1 atributo real continuo [1,...] llamado capital_run_length_average= longitud media de seccuencias ininterrumpidas de letras mayúsculas
- 1 atributo entero continuo [1,...] llamado capital_run_length_longest= longitud de la secuecia de letras mayúsculas ininterrumpida más larga.
- 1 atributo entero continuo [1,...] llamado capital_run_length_total = suma de la longitud de la s secuencias ininterrumpidas de letras mayúsculas / número total de letras mayúsculas en el e-mail
- 1 atributo clase nominal {0,1} llamado spam = denota si el e-mail fue considerado como spam (1) o no (0)

3. En esta tabla no aparecen atributos de valores desconocidos.
4. La **distribución de clases** es la siguiente:

Spam	160	(40%)
No-Spam	240	(60%)

DESCRIPCIÓN DE LA TAREA

1. Preparar los archivos 'spam.names' y 'spam.data' de acuerdo al formato requerido por See5.
2. Entrenar al sistema con estos archivos, interpretando todos los datos relevantes del fichero de salida.
3. Obtener la salida de los e-mails no vistos que aparecen en el fichero 'spam-unseen'.
4. Obtener el conjunto de reglas que describe a los ejemplos de entrenamiento.
5. Modificar el entrenamiento con el fin de introducir la técnica de boosting. Comentar los resultados en cuanto a posibles mejoras del rendimiento.
6. Entrenar mediante la técnica de selección de atributos relevantes (winnowing). Comentar los resultados en cuanto a posibles mejoras del rendimiento.
7. Entrenar mediante la técnica de 10-fold cross-validation. Comentar los resultados en cuanto a posibles mejoras del rendimiento.
8. Entrenar mediante la técnica de validación cruzada (cross-validation), dividiendo para ello el fichero de entrenamiento en una proporción de 60% - 40% para los datos de entrenamiento y test respectivamente.
9. Repetir el experimento varias veces para ver el impacto que tiene la diferente selección aleatoria de ejemplos. Comentar los resultados.
10. Proponer posibles mejoras en cuanto a la recogida de datos, tipos y valores de los mismos,... ¿Te parece un buen método para identificar e-mails spam, o piensas que hay aspectos que no se tienen en cuenta y que podrían ser relevantes?

CRITERIOS DE REALIZACIÓN Y ENTREGA DE LA PRÁCTICA

1. Una guía para la realización de la práctica es el manual del See5. Se debería seguir la descripción que hace este manual como modelo para el comentario de los resultados.
2. La práctica podrá realizarse de forma individual o en grupos de a lo sumo dos alumnos. LA NOTA DE CUALQUIER PRÁCTICA ENTREGADA POR MÁS DE DOS PERSONAS SE REPARTIRÁ A PARTES IGUALES ENTRE LOS COMPONENTES DEL GRUPO.
3. La fecha límite de entrega será el 12 de Abril a las 23 horas, y debe realizarse únicamente por e-mail a la direcciónde l profesor (cmalagon@nebrija.es). En el asunto del mensaje deberá ir la frase *Práctica de Ingeniería del Conocimiento del grupo 5IM1* (o *5IT1* en su caso). La respuesta se dará en un fichero comprimido cuyo nombre debe ser el de los componentes del gupo separados por guión. En dicho fichero irán incluidas la memoria y todo aquel fichero que se obtenga como salida por parte del sistema. Si alguna de estas restricciónes no se cumpliesen serán penalizadas con 0.25 puntos a restar de la nota final de la práctica.
4. No se recogerán prácticas entregadas fuera de fecha o por otro medio distinto de los indicados (disquetes, papel impreso,...)
5. La nota de la práctica es un 70% de la nota final de prácticas.

CRITERIOS GENERALES DE CORRECCIÓN DE LA PRÁCTICA

1. Se valorará la capacidad de analizar los resultados de acuerdo al modelo propuesto (manual del See5), no sólo en cada una de las tareas individuales sino también en cuanto a los mejores o peores resultados compados entre ellas. Una forma de comparación entre los resultados de las diferentes tareas podría hacerse mediante tablas o gráficos.
2. Por supuesto una parte importante de la práctica es la obtención de resultados adecuados.
3. Así mismo se valorarán las aportaciones personales hechas a modo de ampliación, mejoras o aspectos que no se hayan tenido en cuenta.
4. La adecuada presentación de los documentos se da por supuesta. Una mala presentación implica una bajada de nota de hasta un 40%. Dicha presentación no sólo se refiere a la estética de la memoria, sino también, y sobre todo a la aparición de faltas ortográficas o a la mala redacción del texto.