PRÁCTICA DE INGENIERÍA DEL CONOCIMIENTO

CONSTRUCCIÓN DE UN SISTEMA IDENTIFICADOR DE SPAM

Prof. Constantino Malagón Luque
Departamento de Ingeniería Informática

4 de Mayo de 2012

CONSTRUCCIÓN DE UN SISTEMA CLASIFICADOR DE PELICULAS

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 email 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 email 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 secuencias ininterrumpidas de letras mayúsculas

• 1 atributo entero continuo [1,...] llamado capital_run_length_longest= longitud de la secuencia de letras mayúsculas ininterrumpida más larga.
• 1 atributo entero continuo [1,...] llamado capital_run_length_total = suma de la longitud de las 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%)

3. Descripción de la tarea:

1. Preparar los archivos 'spam.names' y 'spam.data' de acuerdo al formato requerido por Weka.

2. Entrenar al sistema con estos archivos usando los siguientes algoritmos:
   a) Arboles de decisión (algoritmo C4.5, llamado en Weka J48)
   b) Algoritmo Naive Bayes.
   c) Algoritmo de los K vecinos más próximos (K Nearest Neighbours)
   d)Algoritmos de combinación de clasificadores: Boosting y Bagging, usando como clasificador individual los árboles de decisión y también con Naive Bayes.

• **NOTA:** Utiliza la técnica de dividir el conjunto de entrenamiento en dos partes (60%-40%) de forma aleatoria para entrenar y testear respectivamente.

3. Explica en un párrafo de 10 líneas como máximo para cada uno en qué consiste cada algoritmo, sin entrar en detalles de cálculos. **NO SE PUEDE COPIAR NADA DE INTERNET. DEBEIS USAR VUESTRAS PROPIAS PALABRAS.**

4.Interpreta todos los datos relevantes del fichero de salida (matriz de confusión, porcentajes de error y clasificación,...) para cada uno de los algoritmos.

5.Haz una tabla comparativa del rendimiento obtenido por cada uno de los algoritmos usados: para ello ten en cuenta como parámetros de comparación los porcentajes de error y el tiempo de realización del experimento. Señala cuál es el algoritmo que creas que es el mejor de ellos.

6. ¿En qué se parecen los mensajes considerados como spam según el árbol de decisión obtenido?

7. Obtener la salida de los e-mails no vistos que aparecen en el fichero 'spam-unseen'.

8. Repetir el experimento varias veces para ver el impacto que tiene la diferente selección aleatoria de ejemplos. Comentar los resultados.

9.Proponer posibles mejoras en cuanto a la recogida de datos, tipos y valores de los mismos, etc. ¿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?

INSTRUCCIONES

1. La extensión de la memoria, en formato opendocument (odt) o pdf, no podrá superar las 25 hojas (siendo la extensión mínima la que se considere oportuna). En dicha memoria deberá ir:
   - Una portada
   - Un índice general
   - Un índice de tablas y figuras
   - Referencias de Internet y/o bibliográficas que hayáis usado para la resolución de la práctica.

AVISO: No se aceptarán memorias que sean entregadas en formato doc (MS Word)

2. La práctica podrá realizarse de forma individual o preferiblemente, en grupos de dos alumnos. En el caso de que se reciba una práctica hecha por tres alumnos se repartirá la nota entre los tres a partes iguales. La fecha límite de entrega será el 25 de Mayo a las 23 horas, y debe realizarse únicamente por e-mail a la dirección del profesor (cmalagon@nebrija.es). En el asunto del mensaje deberá ir la frase *Práctica de Ingeniería del Conocimiento.* La respuesta se dará en un fichero comprimido cuyo nombre debe ser el de los componentes del grupo separados por guión. En dicho fichero irán incluidas la memoria y los ficheros que se crean necesarios. Si alguna de estas restricciones no se cumpliesen serán penalizadas con 0.25 puntos a restar de la nota final de la práctica.

NOTA MUY IMPORTANTE: Cualquier parte de la práctica que se detecte que esté copiada directamente de Internet supondrá una nota automática de 0. En general, en un trabajo de investigación se pueden copiar literalmente pequeños extractos, frases o figuras, pero siempre debe ir entrecomillados o en cursiva, y se debe a su vez citar la fuente de donde se ha extraído.

CRITERIOS GENERALES DE CORRECCIÓN DE LA PRÁCTICA

1. Se valorará la capacidad de analizar los resultados de acuerdo al modelo propuesto, no sólo en cada una de las tareas individuales sino también en cuanto a los mejores o peores resultados comparados 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.