

Reglas de Asociación

Trabajo de evaluación

José Tomás Palma Méndez

Dept. de Ingeniería de la Información y las Comunicaciones. Universidad de Murcia
Contacting author: jtpalma@um.es

Ejercicio 1. El fichero `kddcup99.txt` contiene información sobre diferentes ataques a servidores. Cada línea representa una conexión, entendida ésta como una secuencia de paquetes TCP en un determinado intervalo de tiempo. Cada conexión está etiquetada como normal o con uno de los 19 tipos de ataques considerados. Dichos ataques se agrupan en 4 grandes grupos:

- **DOS**: denegación de servicios.
- **R2L**: acceso no autorizado desde una máquina remota.
- **U2R**: acceso no autorizado a los privilegios del root.
- **Probing**: otra tipo de ataques derivados de técnicas de monitorización o exploración.

En los ficheros que se pueden encontrar junto al archivo con los datos se describe en mayor profundidad la organización de los mismos y la descripción de los atributos.

- 1.a) ¿En qué formato se encuentra el fichero con los datos?
- 1.b) Carga el fichero y comprueba que se ha cargado correctamente.
- 1.c) Reagrupa la información sobre el tipo de ataque para que sólo se tenga información de los cuatro tipos anteriormente citados.

Ejercicio 2. Genera una tabla denominada `items.tab` que contenga dos columnas. La primera columna debe representar distintos valores para el soporte variando de 0.1 a 0.9 con incrementos de 0.1. La segunda columna debe incluir el número de itemsets frecuentes detectados para cada valor de soporte indicado en la primera columna.

- 2.a) ¿Cuál es la longitud máxima de los itemsets que se pueden generar a partir de los datos?
- 2.b) Representa gráficamente dicha tabla.
- 2.c) ¿Cuál es el número máximo de itemsets frecuentes?
- 2.d) ¿Cuál es el mínimo?
- 2.e) Crea un objeto, denominado `attack.trans`, que contenga entre 2000 y 3000 itemsets frecuentes. Elige el valor adecuado para el soporte.
- 2.f) ¿Cuántos de dichos itemsets son maximales? Indica los 3 con el soporte más alto.


- 2.g) ¿Cuántos de dichos itemsets son cerrados? Indica los 3 con el soporte más alto.
- 2.h) ¿Cuántos de dichos itemsets están incluidos en otros itemsets?
- 2.i) ¿Cuántos de dichos itemsets tienen incluidos otros itemsets?

Ejercicio 3. Genera una tabla denominada **rules.tab** que contenga tres columnas. La primera columna debe representar valores desde 0.1 a 0.9 con incrementos de 0.1. La segunda columna debe incluir el número de reglas generadas para cada valor de soporte indicado en la primera columna. La tercera columna deberá contener el número de reglas generadas teniendo en cuenta la confianza indicada en la primera columna. En ambos casos, el parámetro que no se utiliza deberá tener el valor por defecto.

- 3.a) Representa gráficamente dicha tabla.
- 3.b) ¿Cuál es el número máximo de reglas generadas?
- 3.c) ¿Cuál es el mínimo?
- 3.d) Crea un objeto, denominado **attack.rules**, que contenga entre 900000 y 1000000 reglas. Selecciona el valor de soporte adecuado.
- 3.e) Calcula los índice jaccard y el test exacto de Fisher para dichas reglas e incluye dicha información en el conjunto de reglas
- 3.f) ¿Cuántas de dichas reglas son maximales? Indica las 3 con el soporte más alto. Representa todas la reglas maximales con un gráfico de dispersión en el que se muestren en las coordenadas el soporte y el lift y como código de color la confianza. Indica si se puede extraer alguna conclusión del mismo.
- 3.g) ¿Cuántas reglas no redundantes existen? Indica las 3 co el test de Fisher más alto. Representa todas la reglas no redundantes con un gráfico matricial en el que se muestren el soporte y el test de Fisher. Indica si se puede extraer alguna conclusión del mismo.
- 3.h) ¿Cuántas reglas significativas existen? Indica las 3 con el índice jaccard más alto. Representa todas la reglas significativas con un gráfico matricial con reglas agrupadas. Indica si se puede extraer alguna conclusión del mismo.
- 3.i) ¿Cuántas de dichas reglas están incluidas en otras?
- 3.j) ¿Cuántas de dichas reglas incluyen otras?

Ejercicio 4.

- 4.a) Del conjunto de reglas seleccionadas en el apartado anterior ¿Cuántas de dichas reglas nos permiten derivar el tipo de ataque. Muestra por pantalla las 10 reglas con mayor lift junto con las medidas pedidas en el apartado anterior.
- 4.b) Indicar la regla con el soporte más alto que permita determinar qué protocolo está más asociado aun ataque de tipo DOS ¿Qué información nos aporta?.

- 
- 4.c) Representa gráficamente la regla anterior mediante un gráfico de mosaico ¿Qué conclusiones puedes extraer del mismo?
 - 4.d) Del conjunto de reglas seleccionadas en el apartado anterior, elegir la regla con el lift más alto que nos permita determinar qué protocolo y qué servicios están asociados a los ataques los distintos tipos de ataques.
 - 4.e) Indicar cuántas veces se verifica dicha regla para cada uno de los distintos tipos de ataques.