# Trabalho Prático de Datamining para Cybersegurança

Mestrado em Direito e Cybersegurança

# Detectar ligações suspeitas

Ao longo do semestre usou uma base de dados relativa a ligações de rede, que foi usada no KDD em 1999. Há uma base de dados semelhante, ou seja com características de ligações feitas através da internet, registada por um dos sistemas de controlo de uma instituição de referência deste país. Essa base de dados, chamada BNeves.arff tem, para cada ligação feita, os seguintes dados:

**Service Port:** O nº do porto de serviço usado
**Source Port:** O nº do porto de origem da ligação
**Source:** Endereço IP de origem (guardado como um número entre 0 e 4.294.967.296)
**Destination:** Endereço IP de destino (guardado como um número entre 0 e 4.294.967.296)
**Protocol:** Protocolo usado (TCP, UDP, ICMP)[1]

**Acção:** Acção a tomar (Accept, Drop)[2]

A decisão sobre a acção a tomar pode depender de vários factos, e no caso específico desta base de dados, a decisão sobre a acção a tomar não resultou de uma "análise profunda" às ligações (vendo o seu conteúdo, e as suas consequências), mas sim de uma classificação feita por uma dada "firewall". Mesmo assim, pretende-se obter um sistema que classifique cada uma das ligações como normal (e nesse caso deve-se fazer "accept"), ou perigosa (e nesse caso deve-se fazer "drop"). Vamos neste trabalho implementar um sistema desses.

Vamos querer usar este sistema para decidir se devemos ou não aceitar 20 ligações, e as características dessas 20 ligações estão no ficheiro BNeves_novas.arff. De facto, não sabemos qual a acção a tomar para uma dessas ligações, mas de modo a que o ficheiro tenha o mesmo número de colunas que o ficheiro anterior, acrescentámos que devíamos fazer "accept" para todas elas. Se essa não for a acção recomendada pelo classificador, ele indicará um erro, e saberemos que devemos rejeitar essa ligação.

Para tal, siga o guião dado, respondendo às questões feitas.

1. Carregue o conjunto de dados disponíveis para o Weka. Quantas observações tem ?
2. Que tipos de observações tem (normais, e com os diversos tipos de ameaças), e em que quantidade ?

[1] De facto havia ainda IGMP, mas foram acrescentados ao ICMP.

[2] De facto o "Drop" pode corresponder a 4 acções diferentes, mas todas elas implicam intervir: Detect, Reject, Decript, e Drop propriamente dito. No entanto, foram todas agrupadas como apenas "Drop"

3. Obtenha um classificador naive de Bayes para decidir se uma dada ligação é normal ou suspeita. Use 80% dos dados para treino, e 20% para teste.
    a. Qual a taxa de erro do classificador obtido ?
    b. Qual a matriz de confusão do classificador obtido.
    c. Como classifica com este classificador os novos dados ?
4. Obtenha um classificador baseado numa árvore de decisão (por exemplo, C4.5, que aparece no WEKA como J48) para decidir se uma dada ligação é normal ou suspeita. Use 80% dos dados para treino, e 20% para teste.
    a. Qual a taxa de erro do classificador obtido ?
    b. Qual a matriz de confusão do classificador obtido.
    c. Como classifica com este classificador os novos dados ?
5. Obtenha um classificador baseado em vizinhos mais próximos para decidir se uma dada ligação é normal ou suspeita. Use 80% dos dados para treino, e 20% para teste.
    a. Qual a taxa de erro do classificador obtido ?
    b. Qual a matriz de confusão do classificador obtido.
    c. Como classifica com este classificador os novos dados ?
6. Obtenha um classificador baseado em redes neuronais multicamadas para decidir se uma dada ligação é normal ou suspeita. Use 80% dos dados para treino, e 20% para teste.
    a. Qual a taxa de erro do classificador obtido ?
    b. Qual a matriz de confusão do classificador obtido.
    c. Como classifica com este classificador os novos dados ?

Extra:

7.1) Use um classificador diferente, ou escolha uma partição diferente para treino/teste, ou mude os parâmetros de algum dos métodos, e compare com os resultados anteriores.

7.2) Será que consegue obter os mesmos resultados usando menos variáveis ? Justifique a sua resposta, apresentando as taxas de erro obtidas, e as importâncias relativas das diversas variáveis.

Bom trabalho !