UMA METAHEURÍSTICA PARA DE SELEÇÃO DE ATRIBUTOS UTILIZADOS PARA A DETECÇÃO DE ATAQUES CIBERNÉTICOS

Fabio Antero de Pulpa Melo Junior
Instituto Tecnológico de Aeronáutica – ITA
Praça Marechal Eduardo Gomes, 50 – Vila das Acácias, São José dos Campos/SP
Universidade Federal de São Paulo – UNIFESP
Av. Cesare Monsueto Giulio Lattes, 1201 – Eugênio de Melo, São José dos Campos/SP
fabio.antero@unifesp.br / fabioantero@ita.br

Mauri Aparecido de Oliveira
Instituto Tecnológico de Aeronáutica – ITA
Praça Marechal Eduardo Gomes, 50 – Vila das Acácias, São José dos Campos/SP
mauri@ita.br

Lourenço Alves Pereira Júnior
Instituto Tecnológico de Aeronáutica – ITA
Praça Marechal Eduardo Gomes, 50 – Vila das Acácias, São José dos Campos/SP
ljr@ita.br

RESUMO

Os algoritmos de aprendizado de máquina supervisionados, empregados para a detecção de ataques cibernéticos, na maioria das vezes, empregam conjuntos de dados rotulados para treinamento e muitas vezes os atributos necessários ao treinamento são um subconjunto do total de atributos disponível. Utilizar uma quantidade muito grande de atributos para as etapas de treinamento e detecção pode exigir muitos recursos computacionais. Portanto determinar o conjunto mínimo necessário ao treinamento e detecção pode representar uma significativa economia. O objetivo do trabalho era criar uma proposta de mecanismo de seleção de características aplicado ao contexto de detecção de intrusão em redes de computadores que pudesse ser aplicado diretamente sobre os dados sem a necessidade de treinamento e comparação dos resultados obtidos. Para tanto foi utilizada a metaheurística GRASP (Greedy Randomized Adaptive Search Procedure) adaptada para minimizar a distância euclidiana calculada a partir dos vetores dos valores de cada par de atributo. O conjunto de dados utilizado para a realização dos experimentos foi o UNSW-NB15, o qual contém aproximadamente 700.000 registros e 49 atributos. O algoritmo de aprendizado de máquina utilizado foi o k-NN (K-nearest neighbors), por se tratar de um modelo preditivo baseado em distância. O conjunto de dados foi avaliado tanto pela metaheurísitca quanto pelo solver CPLEx para avaliar a precisão dos resultados obtidos pela primeira. Os experimentos foram realizados para conjuntos de 3, 5, 10, 15, 20 e 30 atributos, configurados na forma de um parâmetro. O resultado dos experimentos mostraram que a metaheurísica foi capaz de atingir um valor de distância que apresentou pouca variação. No melhor caso apresentou resultado somente 0,24% inferior ao CPLEx para experimentos realizados com 3 atributos (melhor resultado). No pior caso, obtidos para os experimentos executados com 10 atributos, a variação foi de 7,21%. Uma variação pequena na distância minimizada representa que, na maioria das vezes, os atributos selecionados pelo modelo exato e pela metaheurística são os mesmos. Os atributos que mais contribuem para a minimização da distância euclidiana foram então utilizados para treinar um algoritmo k-NN e os resultados apurados de f1-Score e ROC AUC Score foram obtidos e comparados com os resultados gerados utilizando todos os atributos do conjunto de dados. Com apenas 3 (três) atributos, é possível, para o conjunto de dados utilizado, obter um f1-Score de 0,7758 e uma ROC AUC Score de 0,9603 ao passo que com os 40 (quarenta) atributos os resultados foram respectivamente de 0,9309 e 0,9636, respectivamente.

PALAVRAS CHAVE. Metaheurística, Cibernética e Aprendizado de Máquina.