Data mining approaches for network intrusion detection: from dimensionality reduction to misuse and anomaly detection

Iwan Syarif [1,2], Adam Prugel-Bennett[1], Gary Wills[1]

[1] School of Electronics and Computer Science, University of Southampton, UK
{is1e08,apb,gbw}@ecs.soton.ac.uk
[2] Eletronics Engineering Polytechnics Institute of Surabaya, Indonesia
iwanarif@eepis-its.edu

Abstract

This paper describes the use of data mining techniques to solve three important issues in network intrusion detection problems. The first goal is finding the best dimensionality reduction algorithm which reduces the computational cost while still maintains the accuracy. We implement both feature extraction (Principal Component Analysis and Independent Component Analysis) and feature selection (Genetic Algorithm and Particle Swarm Optimization) techniques for dimensionality reduction. The second goal is finding the best algorithm for misuse detection system to detect known intrusion. We implement four basic machine learning algorithms (Naïve Bayes, Decision Tree, Nearest Neighbour and Rule Induction) and then apply ensemble algorithms such as bagging, boosting and stacking to improve the performance of these four basic algorithms. The third goal is finding the best clustering algorithms to detect network anomalies which contains unknown intrusion. We analyze and compare the performance of four unsupervised clustering algorithms (k-Means, k-Medoids, EM clustering and distance-based outlier detection) in terms of accuracy and false positives.

Our experiment shows that the Nearest Neighbour (NN) classifier when implemented with Particle Swarm Optimization (PSO) as an attribute selection algorithm, achieved the best performance, which is 99.71% accuracy and 0.27% false positive. The misuse detection technique achieves a very good performance with more than 99% accuracy when detecting known intrusion but it fails to accurately detect data set with a large number of unknown intrusions where the highest accuracy is only 63.97%. In contrast, the anomaly detection approach shows promising results where the distance-based outlier detection method outperforms the other three clustering algorithms with the accuracy of 80.15%, followed by EM clustering (78.06%), k-Medoids (76.71%), improved k-Means (65.40%) and k-Means (57.81%).

Keywords: intrusion detection system, anomaly detection, misuse detection, feature selection, clustering, ensemble classifiers