Contribution ID: **122** Type: **Poster Presentation**

False Positive Reduction in Intrusion Detection System Based on Transfer Learning

In network intrusion detection systems, alert logs generated by intrusion detection devices contain a large number of false positive alert logs, which seriously impair the accuracy of security incident analysis. Thus, filtering false positive alert logs is of great significance. The essence of false positive alert filtering is a classification task: each alert log is labeled to indicate whether it is a false positive, and supervised learning can accurately identify such false positive alert logs. However, the availability of logs with the aforementioned labels is very limited, resulting in unsatisfactory performance of traditional supervised learning methods. To solve this problem, this paper proposes a false alert filtering method based on transfer learning, which aims to realize intelligent discrimination of alert logs by utilizing a small number of logs with the aforementioned labels and a large number of logs without such labels. In this method, the aforementioned labels are defined as source domain labels, and all logs are assigned a target domain label indicating whether the log possesses a source domain label. Subsequently, a Domain-Adversarial Neural Network (DANN) is introduced, comprising a feature extractor, a label predictor, and a domain classifier. The feature extractor conducts feature extraction on the original logs, the label predictor identifies false positive logs, and the domain classifier determines whether a log has a source domain label based on the log's target domain label and its extracted features. The Gradient Reversal Layer (GRL) ensures that the feature distributions of logs with source domain labels and those without become similar—rendering the domain classifier unable to distinguish between the two—ultimately achieving accurate discrimination of alert logs.

Primary authors: [redacted]; [redacted]; [redacted]; [redacted]

Presenter: [redacted]

Track Classification: Track 7: Network, Security, Infrastructure & Operations