

Title	人と協調してネットワーク運用を支援する機械学習に関する研究
Author(s)	川口, 英俊
Citation	
Issue Date	2023-03
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/18426
Rights	
Description	Supervisor: 岡田 将吾, 先端科学技術研究科, 博士

Abstract

Communication networks have become indispensable to people's lives. In this age of smartphones, the Internet can be regarded as an infrastructure for daily life and electricity, water, and gas. Companies that provide such communication services to users struggle daily to ensure the stable communication networks.

Network operations experts are needed to ensure the stable communication networks. However, with new technologies such as the Internet of things (IoT) and 5G, the burden on the experts who manage them continues to increase. The work of network operations experts is diverse and includes many decision-making tasks.

We focus on the intrusion detection and prevention system (IDPS) signature classification task in network operations. IDPSs monitor network systems and take actions such as logging, notification, and blocking when malicious communications are detected. This dissertation focuses on a type of IDPS that performs detection based on pattern files of malicious communications such as signatures. The signatures are distributed periodically by the IDPS developers, similar to a subscription service. Network operations experts determine the importance level ("low", "medium", or "high") of each signature to set IDPS actions. For example, if the importance of a signature is "high", the action is "blocking"; if it is low, the action is "logging".

Determining the importance level of a signature for setting it up in the IDPS is a burden for the expert. While expert-designed if-then rule scripts can automatically determine some signatures, the remaining signatures must be determined manually by experts based on elements in the signatures, articles on the Internet, and their own experience. This manual decision-making process takes some time. In addition to time consuming issues, it takes sufficient knowledge and experience to determine the importance level, and there are not many experts who have such knowledge and experience. In other words, the cost of hiring experts is high. Therefore, determining the importance of a signature, which requires an expert's time, is also a significant cost to the network company.

IDPS signature classification is an important task, but there has yet to be research to automate it. We must recognize the seriousness of classification errors that can result from automation, as is the case in the medical field. Signature misclassification leads to IDPS misconfiguration. Misconfiguration of IDPS can cause security incidents, such as missing malicious communications and false interceptions of regular communications. Security incidents should be avoided because they damage public trust in the organization operating the network system. Hence, it is not practical to automate all signature classifications. In order to automate

classification while reducing classification errors, a framework for efficient classification in cooperation with humans, such as checking with experts as necessary, is required.

This dissertation aims to formulate IDPS signature classification as a machine learning problem for the first time and to build and evaluate a system that cooperates with experts to classify signatures. To achieve this goal, we addressed three problems.

First, there are no publicly available datasets for machine learning signature classification. In other words, they need to prepare for the prerequisites of the research. Several reasons make signature datasets difficult to make publicly available: Many of the signatures are distributed by IDPS developers, but they cannot be redistributed under license; Publishing the signatures and their labels may lead to the outside world guessing about the sensitive information of the IDPS configuration. We collected the three datasets used in this research in cooperation with several experts in actual network operations organizations. These are real datasets consisting of signatures that experts actually set in the IDPS. Experts classify some signatures by predefined if-then rules. An if-then rule returns a label of “low”, “medium”, “high”, or “unknown” importance based on keyword matching of the elements in the signature. Two datasets, the automatically annotated dataset (AAD) and the manually annotated dataset (MAD), were collected. AAD consists of 4,465 signatures automatically labeled by expert-designed if-then rule scripts. MAD consists of 1,300 signatures that could not be classified by the if-then rule scripts and were manually labeled by the experts. Next, we collected a time-series manually annotated dataset (TMAD) consisting of 7,577 signatures that were manually labeled and time-stamped with the date and time of distribution. Both labels of signatures were determined after consultation with several experts. This research is based on these three datasets.

Second, to classify IDPS signatures by machine learning, it is necessary to search for an effective feature extraction method. We propose three features based on the expert’s knowledge, with interpretability to clarify the expert’s criteria. We first design two types of features, symbolic features (SFs) and keyword features (KFs), which are used in keyword matching for the if-then rules. Next, we design web information and message features (WMFs) to capture the properties of signatures that do not match the if-then rules. The WMFs are extracted as term frequency-inverse document frequency (TF-IDF) features of the message text in the signatures. The features are obtained by web scraping from the referenced external attack identification systems described in the signature. The effectiveness of the proposed features is evaluated in experiments with AAD and MAD. In the experiment, the classification models with proposed features are evaluated from two perspectives: classification accuracy and reject option (RO) performance. In

both cases, the combined SFs and WMFs performed better than the combined SFs and KFs. We also show that using an ensemble of neural networks (deep ensembles; DE) improves the performance of the RO. An analysis shows that experts refer to natural-language elements in the signatures and information from external information systems on the Internet.

Third, if a fully automated machine learning model replaces the IDPS signature classification task, there is a risk of missing critical classification errors. It is also necessary to entrust experts with decisions that have a high risk of error by signature classification models. Therefore, it is important to establish a method for humans and the system to cooperate in setting up and classifying data. In addition, to actually use machine learning, it is necessary to cope with high annotation costs and domain shifts caused by signatures created to keep up with new cyber attacks. In this dissertation, we propose a system based on active learning in cooperation with experts to overcome three challenges: (a) security incidents caused by classification errors, (b) high annotation costs, and (c) classification accuracy decrease due to domain shifts. The proposed system includes an IDPS signature classification model and periodically classifies the received signatures in cooperation with an expert. The uncertainty sampling is used as an acquisition function to preferentially transfer signatures with a high risk of misclassification to the expert. The signatures are sorted by uncertainty sampling; some are transferred to experts, and the rest are automatically classified. The experts classify the transferred signatures and add them to the training dataset, and the classification model is retrained. After training, the new signatures that have not yet been labeled are classified. The proposed system executes this workflow each time it receives signatures. Uncertainty estimation methods in deep learning, such as Monte Carlo dropout (MC-Dropout) and DE, are also incorporated to identify signatures at high risk of misclassification accurately. Experiments are conducted on the TMAD to evaluate the proposed system in a simulation case. An analysis is then performed by comparing several variants of the proposed system. The results show that the system with MC-Dropout performs best. We also show that this variation has two effects: it transfers more samples with “medium” importance to the experts and mitigates imbalances in the training dataset.

As described above, in this dissertation, we collected IDPS signature datasets that are difficult to make public, and proposed features for machine learning classification of IDPS signatures and an active learning-based system to cooperate with experts. The proposed system enables accurate identification of IDPS signatures and contributes to reducing fatal classification errors, which are problematic in practical applications. Analysis using the proposed features identifies the elements in the signatures that are important to experts when classifying signatures. Identifying the important factors to experts can provide helpful information for other

machine learning and non-machine learning approaches to signature classification. The proposed system procedures are widely applicable not only to signatures. There are other tasks in network operations where data are generated periodically and classified by experts. For example, software vulnerability information, such as common vulnerabilities and exposures (CVE), is released periodically, and experts may decide whether to classify this information as necessary. In this dissertation, task sharing is considered collaboration, but interaction with machine learning systems and education of novices using them are also examples of collaboration. The realization of such collaborations is future works for machine learning technology to support network operations. We hope that the ideas and evaluation results in this dissertation will help solve signature classification problems as well as other tasks.

Keywords: machine learning, IDPS, signature, active learning, uncertainty estimation, reject option.