BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2023
http://excel.fit.vutbr.cz

Information Fusion for Classification of Network Devices

Bc. Ondřej Sedláček*

Abstract
An essential aspect of cybersecurity management is maintaining knowledge of the assets in the protected network. Automated asset discovery and classification can be done using various methods, differing in reliability and the provided type of information. Therefore, deploying multiple methods and combining their results is usually needed – but this is a nontrivial task. We present a two-layer data fusion approach that can effectively fuse multiple heterogeneous and unreliable sources of information about a network device to classify it. The solution is based on a combination of expert-written conditions, machine learning from small amounts of data, and the Dempster-Shafer theory of evidence. Experiments show that our method is on par with the best ML-based methods in classification accuracy but with the advantage of better interpretability and robustness against some types of input data imprecisions that can occur in practice.

*xsedla1o@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

The importance of maintaining knowledge of assets in the monitored network is often neglected in cybersecurity management. With complex and dynamic networks, more than relying on manual documentation is required; a system for automatic asset discovery and classification is needed.

Traditional methods of obtaining information about connected devices involve periodic scanning of the internal network or obtaining useful information about devices using passive network traffic monitoring [1, 2, 3, 4]. Combining active and passive approaches gives the most comprehensive and up-to-date view of network assets; however, combining data from multiple heterogeneous and unreliable sources is nontrivial.

We present a new data fusion method that effectively combines data from different asset discovery tools to solve the issues encountered in practice. The approach combines an expert-defined condition layer with a machine learning classifier based on the Dempster-Shafer theory of evidence [5, 6]. Several machine learning models are experimentally evaluated for classification performance, interpretability, and robustness against input data flaws.

2. Solution overview

This section demonstrates our solution with an actual deployment example from the Asset Discovery, Classification and Tagging (ADiCT) project. The ADiCT project aims to create a knowledge base about the monitored network using passive monitoring methods. However, the various results from these methods pose a problem for determining a singular output, as shown in the left half of Figure 1.

This leads us to our problem statement: We aim to classify entities in a monitored network in aspects such as the operating system and device type, with input sources in different formats, classifications, and taxonomies. Our solution must be configurable and interpretable while significantly improving output accuracy. Some data sources may support each other, thus increasing overall confidence, while others may contradict. Another issue is that not all possible data sources provide data for most objects.

The right half of Figure 1 illustrates our solution using the Dempster-Shafer theory model, where the input data is normalized using expert-defined conditions and transformed into belief functions. These belief functions are merged using Dempster's Rule of Combination, noted as $\oplus$, to produce a singular belief function as the model output.

3. Information Fusion Model

Our model consists of two layers: the condition layer and the combination layer. The condition layer, shown in Figure 2, consists of expert-written conditions that may or may not be satisfied by the input vector, represented as 1 or 0, respectively. The input of the condition layer is a vector of values describing the device, and the output is a feature vector encoding all relevant information.

The combination layer takes the binary feature vector as input and predicts the most probable class of the device in a given taxonomy using a trained classifier. Based on our experiments, we have selected an interpretable classifier based on the Dempster-Shafer theory. As a result, it is easy to derive which conditions, and therefore which input attributes and their values, had a decisive influence on the final classification. The classifier's ability to explain the results makes users trust the system, making it a more helpful tool [7, 8]. The complete information fusion model is illustrated in Figure 3.

4. Evaluation

We have conducted a series of experiments to evaluate our approach. First, we used a real dataset from our network to test the solution in our use case. In this case, we focused on classifying the operating system running on each device. Then, we used a series of generated datasets to further explore our approach's properties in different conditions.

In both cases, we tested multiple classifiers in place of the combination layer to see which achieves the best results. These include classifiers based on the Dempster-Shafer theory, where we note our approach **D-S1** and an approach inspired by [9, 10] **D-S2**. Dempster-Schafer Gradient Descent (**DSGD**) is a re-implementation of an approach described by Peñafiel et al. [7]. Weighted majority voting (**WMV**) is a simple approach where each condition gets a 'vote', and all votes sum up to the final classification. Three additional machine learning models are evaluated: Decision Tree (**DT**), Random Forest (**RF**) and AdaBoost (**AB**). For reference, we also include an Oracle (**ORA**) classifier, representing the ideal fusion method. The Oracle classifier has access to the annotation data and correctly classifies the input vector as long as at least one of the input data sources correctly classifies it.

The real network dataset, which was collected over **93** days, contains **5532** samples (classifiable sessions) of **674** unique IP addresses. The dataset was balanced by excluding samples from the Linux class, as it is over-represented in a 3:1 ratio to the second most represented class in the captured data. The dataset shows a relatively low amount of data source overlap, e.g. data are available from only one source in 63% of sessions and two sources in 30% of sessions.

Figure 4 shows the results achieved by using each of the tested classifiers in terms of accuracy and F1-score: The DT and RF classifiers show the best results, followed by the DSGD implementation. The ensemble classifier AB shows results comparable to D-S1 and D-S2. WMV has both the worst accuracy and F1-Score, which is not unexpected considering the classifier's simplicity.

Overall, the results show that all tested models are similar in performance. In this case, the DT classifier outperforms the others by a small margin. These results can be explained by the simplicity of the scenario, given by the relatively low number of concurrent data sources. The following section shows that DT does not perform so well in other scenarios.

To allow us to explore the behavior of our model in different circumstances, we generated a series of datasets. We explore the influences of parameters such as the number of data sources and classes. On top of that, we simulate different scenarios of setting the accuracy of data sources. The scenarios include having a uniform or variable accuracy distribution for data sources, adding various faults that realistically distort data and random sampling, a combination of all previously described scenarios.

Figure 5 shows the average scores of compared methods over all generated dataset categories. Methods are sorted by the *Random* column, which are the F1-score averages over 20 randomly selected combinations of experiment parameters. Again, we can see that all methods give similar results. The D-S1 classifier leads by up to 2 percent depending on the scenario, and DSGD follows second. The D-S2 and WMV classifiers show the best results in uniform accuracy scenarios or scenarios with mild variability but lose their advantage after adding more severe faults to the inputs. Results of the RF, AB, and DSGD classifiers show the opposite: better performance in scenarios with lower input accuracies. The DT classifier has shown the worst results in this more complex comparison, especially in experiments with fault resilience. In conclusion, the D-S1 classifier is the most versatile solution that, on average, adapted best to all the examined scenarios.

References

[1] Richard Lippmann, David Fried, Keith Piwowarski, and William Streilein. Passive operating system identification from TCP/IP packet headers. In *Workshop on Data Mining for Computer Security*, volume 40, 2003.

[2] Takashi Matsunaka, Akira Yamada, and Ayumu Kubota. Passive OS Fingerprinting by DNS Traffic Analysis. In *2013 IEEE 27th International Conference on Advanced Information Networking and Applications (AINA)*, pages 243–250, 2013.

[3] Mustafizur R Shahid, Gregory Blanc, Zonghua Zhang, and Hervé Debar. IoT devices recognition through network traffic analysis. In *2018 IEEE international conference on big data (big data)*, pages 5187–5192. IEEE, 2018.

[4] Martin Husák, Milan Cermak, Tomas Jirsik, and Pavel Celeda. HTTPS traffic analysis and client identification using passive SSL/TLS fingerprinting. *EURASIP Journal on Information Security*, 2016, 02 2016.

[5] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Institute of Mathematical Statistics*, 38(2):325–339, April 1967.

[6] Glenn Shafer. *A mathematical theory of evidence*. Princeton university press, 1976.

[7] Sergio Peñafiel, Nelson Baloian, Horacio Sanson, and José A Pino. Applying Dempster–Shafer theory for developing a flexible, accurate and interpretable classifier. *Expert Systems with Applications*, 148:113262, 2020.

[8] Swetha Hariharan, Anusha Velicheti, A.S. Anagha, Ciza Thomas, and N. Balakrishnan. Explainable Artificial Intelligence in Cybersecurity: A Brief Review. In *2021 4th International Conference on Security and Privacy (ISEA-ISAP)*, 2021.

[9] Ahmed Al-Ani and Mohamed Deriche. A new technique for combining multiple classifiers using the Dempster-Shafer theory of evidence. *Journal of Artificial Intelligence Research*, 17:333–361, 2002.

[10] Alberto Dainotti, Antonio Pescapé, and Carlo Sansone. Early classification of network traffic through multi-classification. In *International Workshop on Traffic Monitoring and Analysis*, pages 122–135. Springer, 2011.