Information Fusion for Classification of Network Devices

author
Bc. Ondřej Sedláček

supervisor
Ing. Martin Žádník Ph.D.

technical supervisor
Ing. Václav Bartoš Ph.D.

Motivation

An important aspect of network monitoring is to keep track of devices connected into the monitored network. For this purpose, various methods of automatic asset discovery and classification are being used. Therefore, deploying multiple methods and combining their results is usually needed — but this is a non-trivial 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.

Solution overview

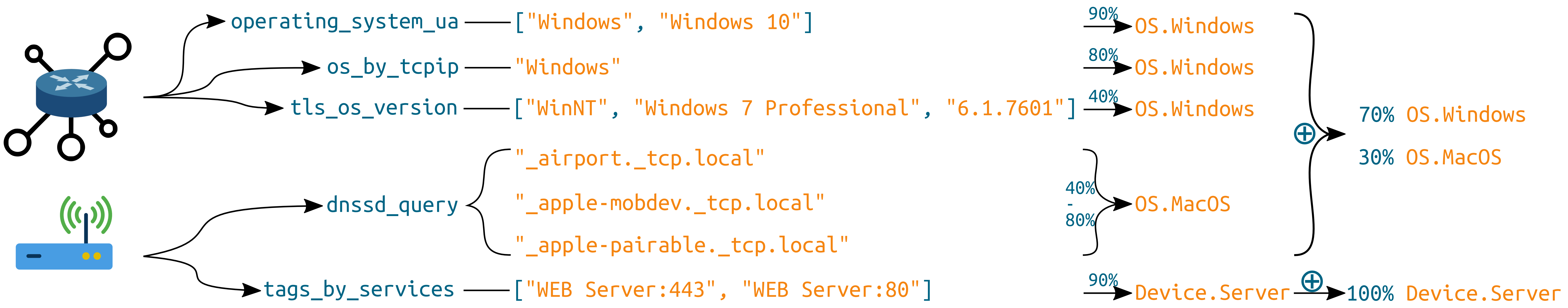

operating_system_ua — ["Windows", "Windows 10"]
90% OS.Windows
os_by_tcpip — "Windows"
80% OS.Windows
tls_os_version — ["WinNT", "Windows 7 Professional", "6.1.7601"]
40% OS.Windows
dnssd_query
"_airport._tcp.local"
"_apple-mobdev._tcp.local"
"_apple-pairable._tcp.local"
40% - 80% OS.MacOS
⊕
70% OS.Windows
30% OS.MacOS
tags_by_services — ["WEB Server:443", "WEB Server:80"]
90% Device.Server
⊕
100% Device.Server

Figure 1: Solution showcase

- Solution showcase on specific use case of the **ADiCT project**
  - ADICT is focused on development of new passive monitoring methods
- **Goal**: Fusion of all available device data

Input: Attributes from **various classifiers** and monitoring methods

- **Diverse input data**: different types, formats, levels of detail

Output: Fused classification for given entity

- Classified aspects: **operating system, device type**

Key features

- Fusion method **robust against missing inputs**
- Solution **maintains level of detail** of individual sources
- **Configurability**; Open to new data sources
- Leads to significant **improvement in output accuracy** (compared to individual data sources)
- **Interpretable**, explainable

Information Fusion Model

- **Diverse data sources** require **data normalization**
- **Condition layer** — expert defined **conditions** used to normalize attributes to **a binary vector**.
- **Expert knowledge is not enough**; Determining condition output directy gives bad results - only select relevant attribute values

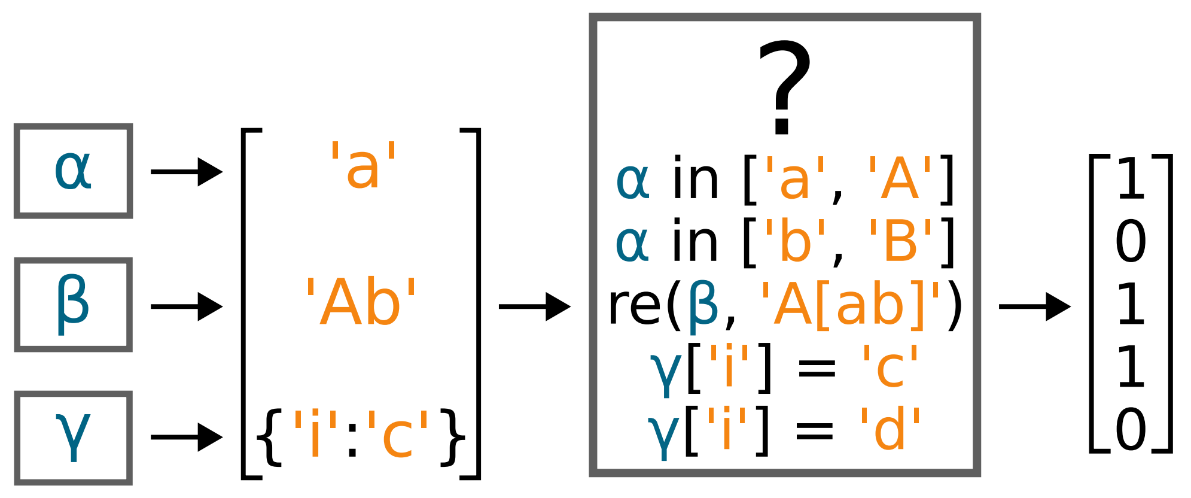


Figure 2: Condition layer transformation. Values from data sources α, β and γ are transformed into a vector encoding all relevant information the values contain.

- **Combination layer** — outputs the most probable class
- Any **interpretable classifier** can be used

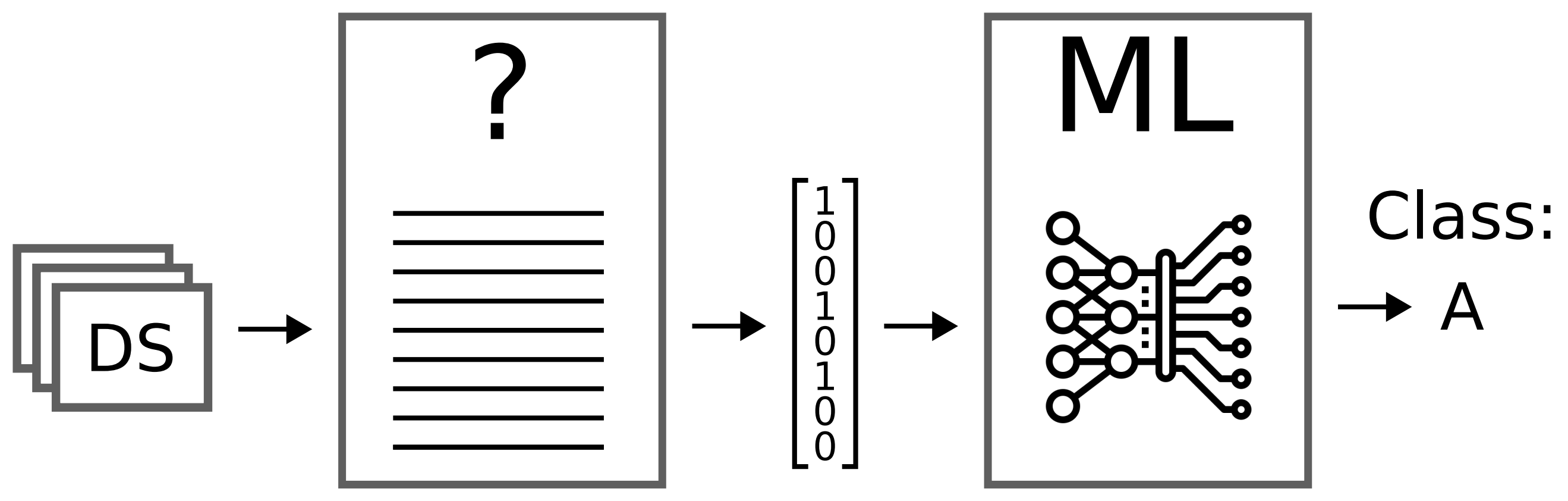


Figure 3: Proposed information fusion model. Values from an arbitrary number of data sources are first transformed by the condition layer. The combination layer determines the output class.

Tested classifier acronyms: **ORA**: Oracle, **D-S**: Dempster-Shafer, **DSGD**: Dempster-Shafer Gradient Descent, **AB**: AdaBoost, **WMV**: Weighted Majority Voting, **RF**: Random Forest, **DT**: Decision Tree

Evaluation

- **Real network dataset**: **5532** samples (classifiable sessions) of **674** unique IP addresses, collected over **93 days**
- Tested **various interpretable classifiers**

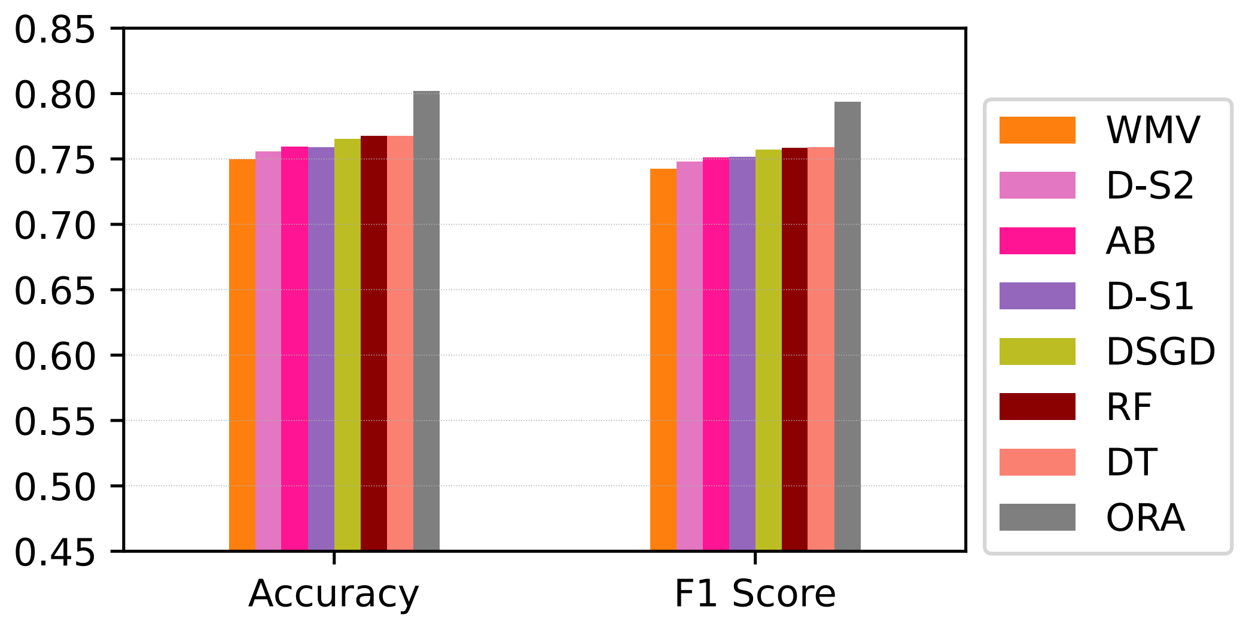
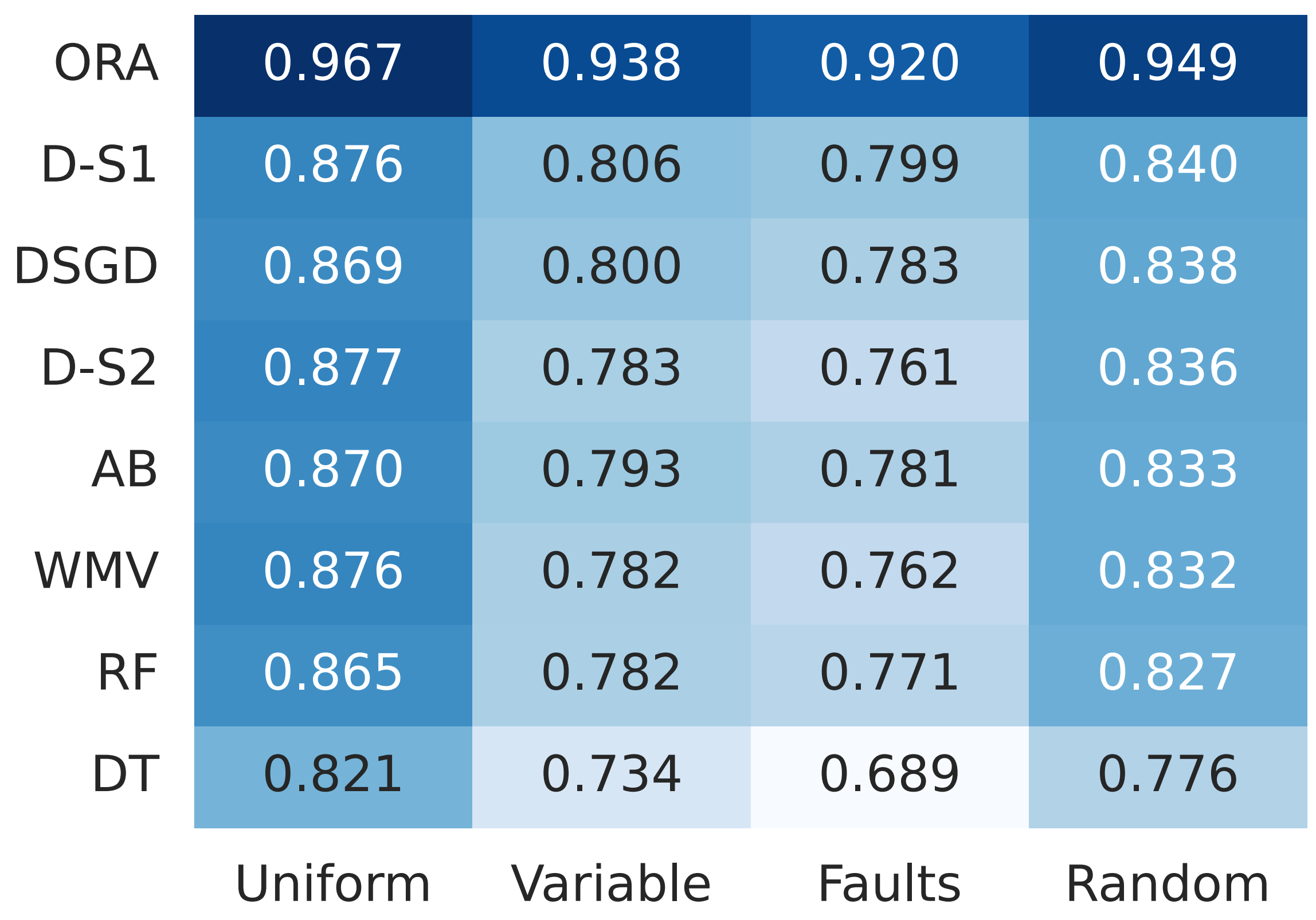


Figure 4: Fusion methods accuracy and F1-Score.

- **Little data** source **overlap** in real dataset — one possible scenario
- Testing under **different conditions**: series of generated datasets
  - Influence of **number of sources**, **variability**, source **faults**



	Uniform	Variable	Faults	Random
ORA	0.967	0.938	0.920	0.949
D-S1	0.876	0.806	0.799	0.840
DSGD	0.869	0.800	0.783	0.838
D-S2	0.877	0.783	0.761	0.836
AB	0.870	0.793	0.781	0.833
WMV	0.876	0.782	0.762	0.832
RF	0.865	0.782	0.771	0.827
DT	0.821	0.734	0.689	0.776

Figure 5: Individual method scores. Averaged over all experiment categories

BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY
cesnet
Excel@FIT 2023