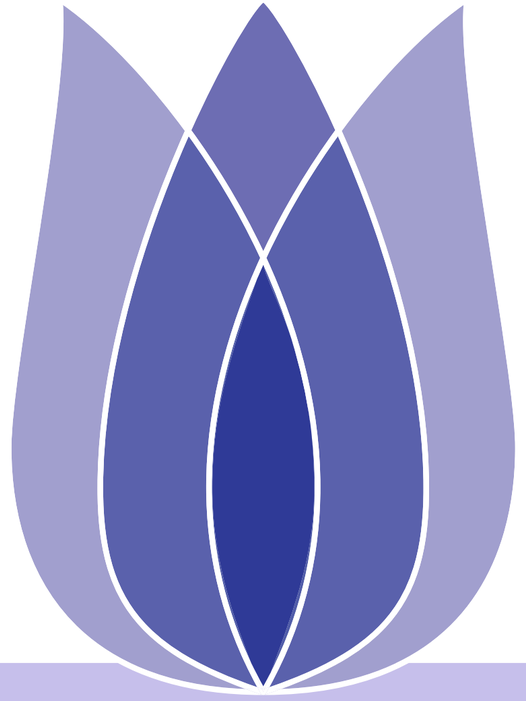
CDMC2020 Task Presentation

Rongxin Xu, Xiaoyan Wang

2020-10-04



Overview

Problem Definition

Task1

Current Work and Challenges

Current Work

Problem Definition

Task1

Defn Predict the malware category based on the given data and software hash value.

- Data imbalance
- How to use hash value

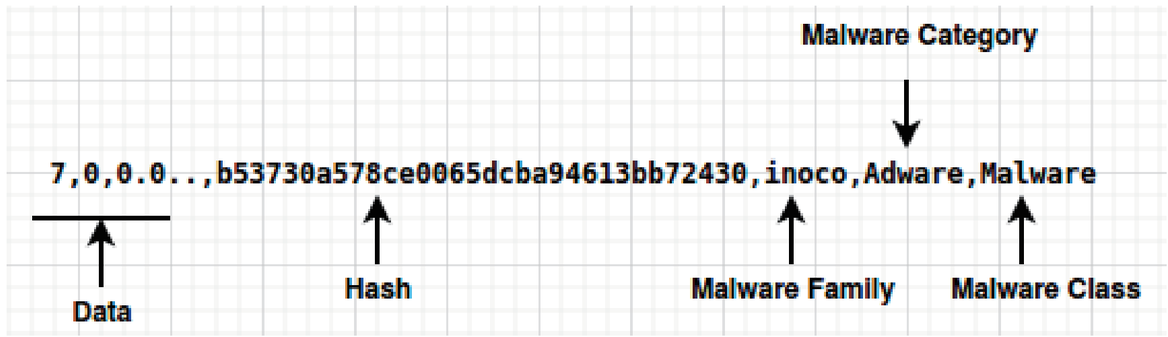


Figure 1: The data structure of the training file

Task2

Defn Predict the category of malware based on a given byte sequence.

Current Work and Challenges

Current Work

For task 1, the following work has been completed.

- labelencoder
- Dimensionality reduction
- Comprehensive sampling(SMOTE + Tomek links)
- Cross-validation
- Base model(XGBoost, KNN, SVM...)
- Model stacking

For task 2, Related work needs to be carried out.

Current Work

Here is the ensemble architecture:

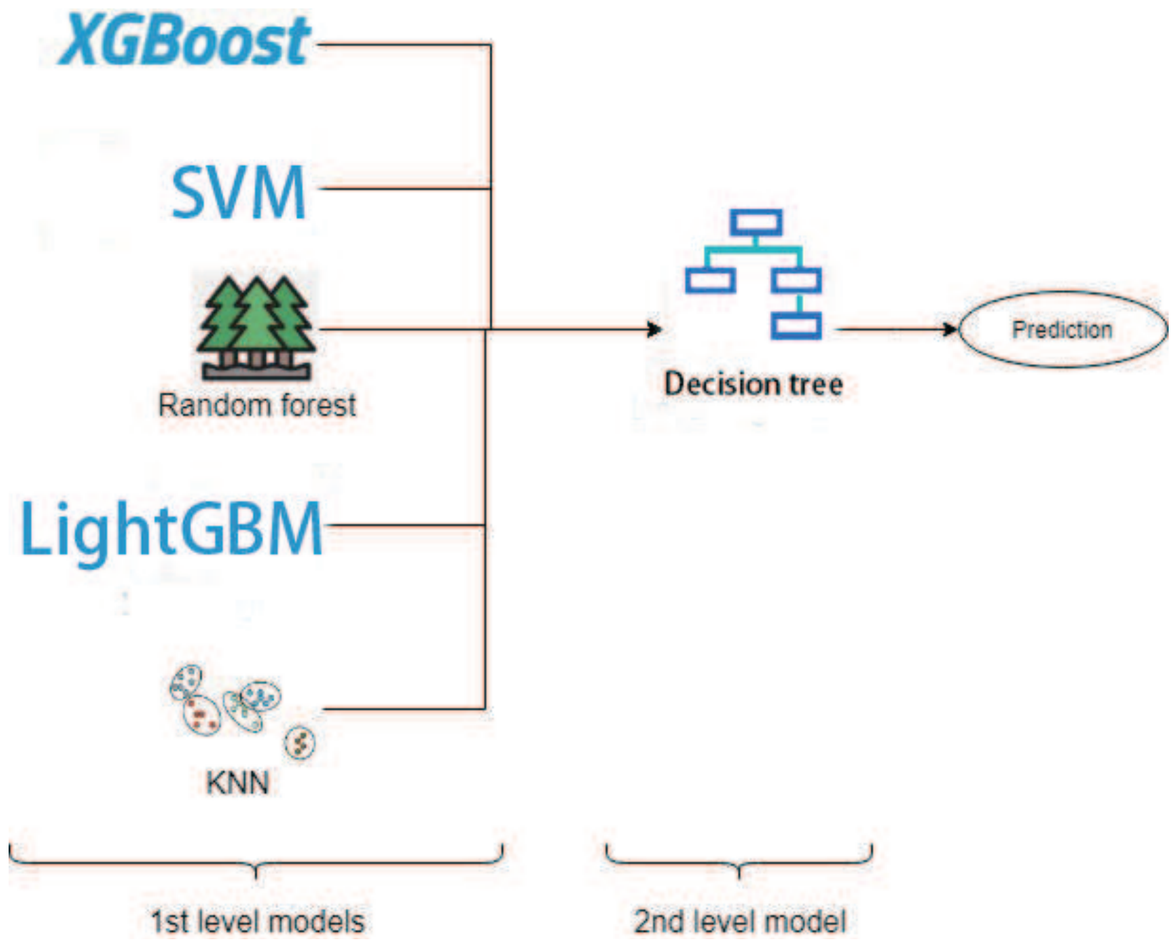


Figure 2: ensemble architecture

Challenges

- How to use hash values and byte sequences.
- Is the NLP model effective?
- Report

End

Thank you!