The Construction of Rules for EMBER Dataset with the Help of a Decision Tree Model

Ján Mojžiš[1][0000-0002-2196-2271]

[1] Institute of Informatics, SAS, Bratislava, Slovakia
jan.mojzis@savba.sk

Abstract. Ontology is traditionally used in concept learning. Decision tree models were used before in the process of ontology creation, in applications of recommender systems or for prediction in a manufacturing network. Decision tree model in comparison to neural networks can generate interpretable rules based on the features located in dataset. In this paper, a decision tree model, which was created with the Waikato Environment for Knowledge Analysis (WEKA) application, is used to generate rules from EMBER dataset. In the result, three rules are generated, covering 52.6% of all malware samples presented in the dataset.

Keywords: Ontology, decision tree, EMBER, malware.

1 Introduction

An ontology is a formalized, structured and machine-readable representation of data. A learning software can be used to process it and to discover the rules or concepts contained in the data. In concept learning, we use the software learner to discover concepts or rules that should describe the data and relations between entries. A traditional machine-learning practices can be used to evaluate the features in the ontology data. Decision trees can help with rules or concepts constructions, finding appropriate relations and features. In this paper we apply J48 decision tree approach to ontological data, discovering three rules, covering more than 50% of all malware entries, contained in the dataset.

2 Related works

Zalan et al. [3] propose a predictive model, which assist in the allocation of newly received orders in a manufacturing network. The methodology presents the mapping of a PROSA (Product-Resource-Orcer-Staff Architecture) based ontology on a decision tree, created with the Waikato Environment for Knowledge Analysis (WEKA) application [1]. A decision tree algorithm was mapped into ontology with the help of Semantic Web Rule Language (SWRL). The SWRL rules were extracted from the

decision tree model with the help of a MATLAB program. In the result a model gave 60.4% prediction accuracy.

Bouza et al. [2] propose SemTree for the use in the recommender system. SemTree is an ontology-based decision tree learner, that use a reasoner and an ontology to semantically generalize item features to improve the effectiveness of the decision tree built. He evaluates SemTree with J48 and outperforms it [1].

In the objective of this paper the ontology is already given, and I search for the rules which should help to discriminate malware entries from benign entries.

3 Dataset and ML model used

I have used Weka's J48 Tree model in Weka [1]. The dataset EMBER [4] is a collection of statically extracted features from approximately 1.1 million benign/malicious Windows executables. I have reduced the original dataset to 500 samples of malware and 500 samples of benign-type records. Features were filtered with the help of CFS [5] feature selector, available in Weka.

4 Results

The tree model generated the following rules:

1. #pe_high_entropy > 1, data <= 7.246707 (137 of which 6 is false)
2. #pe_high_entropy <= 1, code > 6.522143, #pe_registry > 0 (54 of which 2 are false)
3. #pe_high_entropy <= 1, code <= 6.522143, #pe_dll <= 0, data > 3.799375, data <= 4.083352, rsrc <= 5.614608, rdata > 5.0436 (72 of which 1 is false)

The rules 1. and 2. can be seen in visualized tree in Fig. 1.

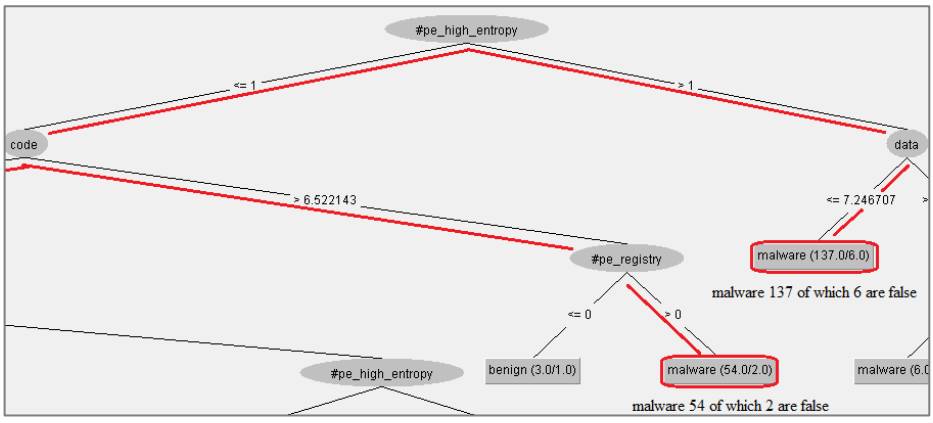

Fig. 1. An excerption of the generated J48 Tree from the sample dataset.

5 Conclusions

In this short paper the application of decision tree on EMBER dataset is presented. Three rules were generated which together are covering 263 out of 500 malware samples, contained in the reduced dataset. Reduced dataset proved that there are rules discriminating malware and benign samples.

References

1. B. Sarwar, G. Karypis, J. Konstan, and J. Riedl.: Item-based collaborative filtering recommendation algorithms. In Proceedings of the 10th international conference on World Wide Web, ACM, Hong Kong (2001).
2. Bouza, A., Reif, G., Bernstein, A., Gall, H.: SemTree: ontology-based decision tree algorithm for recommender systems. In: International Semantic Web Conference, Karlsruhe, Germany (2008).
3. Khan, Z.M.A., Saeidlou, S. and Saadat, M.: Ontology-based decision tree model for prediction in a manufacturing network. Production & Manufacturing Research *7*(1), 335-349 (2019).
4. Anderson, H. S. and Roth, P.: Ember: an open dataset for training static pe malware machine learning models. arXiv preprint arXiv:1804.04637.
5. Hall, M.A., 1998.: Correlation-based feature subset selection for machine learning. Thesis submitted in partial fulfillment of the requirements of the degree of Doctor of Philosophy at the University of Waikato (1998).