CDMC 2020 TASK DESCRIPTION REPORT

RONGXIN XU, XIAOYAN WANG

Abstract. Based on the byte sequences collected at the entry points of ELF files as discriminant features and the malware families of the programs as training labels, the participants are required to perform a classification task to predict the malware families of the test samples.

Contents

Date: 2020-10-25.
2020 *Mathematics Subject Classification.* Malware detection.
Key words and phrases. Random forest, PCA, ...

1. Task 1

1.1. **Problem Analysis.** This is a typical classification problem, and we can certainly transform it into other problems, such as image detection, but we believe that the classic classification problem has better operability. But we face some difficulties. On the one hand, this is a data set with tens of thousands of samples and thousands of dimensions. At the same time, a large number of attributes are almost filled with 0. These redundant attributes have an adverse effect on the training of the model. On the other hand, the data set is in a severely unbalanced state, which poses a challenge to the accuracy of the classifier.

1.2. **Our work.**

We have used some basic but effective methods to solve the above problems, the main methods used are as follows.

PCA: Considering the consistency of the training set and the test set, we did not choose a manifold learning method such as T-SNE to reduce dimensionality. We used the PCA method to reduce the dimensionality to the most effective number.

SMOTE and Tomek Links: The data set is in a severely unbalanced state, and the simplest and most effective method is to resample. In order to maximize the balance of the negative effects of different resampling methods on the model, we decided to use a comprehensive sampling method: SMOTE + Tomek Links.

Random forest: We observed the performance of many classification models using cross-validation, and finally decided to use random forest.

2. Task 2

2.1. **Problem Analysis.** This is also a classification problem, the difference is that the data type of the attributes of the data set is string type, which is not accepted by machine learning models. We need to find a way to map strings to numbers.

2.2. **Our work.**

We also found a simple and easy way to map a string to a number, at the same time, the data set has more serious imbalance than Task 1. In order to provide more information to the model, we also did some additional work.

ASCII: Since the length of the string is fixed, we consider converting each letter to its corresponding ASCII value. In this way, we have constructed 2732 attributes. Each attribute has a number corresponding to the letter at the same position in the string.

One-Hot Encoding: In order to make full use of the information in the data set, we have one-hot encoding the "cpu" attribute.

ensemble learning: We divide the data set into several parts according to the number of samples, the number of samples in each part is at a similar level, and then the corresponding first-level classifiers are trained respectively, and finally the predicted probabilities of the first-level classifiers are integrated .

References

List of Todos

(A. 1) School of Business Administration,, Hunan University, Changsha 410012, China
Email address, A. 1: rongxin_xu@163.com