共同研究報告

Solution for The CDMC 2017

Yuki Maruno, Ayumi Hirao, Mayu Nishimoto,
Midori Sakai, Marie Ohki

Abstract

The CDMC 2017 is a competition focusing on real-world problems regarding cybersecurity. We took part in this competition and our team was the first place winner. In this paper, we describe how we solved the following tasks with the provided dataset. We used the Random Forest classifier for all the tasks with the hyperparameter optimization and the feature selection. Experiments showed that our proposed method can obtain an accuracy more than 90% without high computational costs.

Key words：Data Mining Competition, APK Malware, Incident Detection, Fraud Detection, Random Forest

1　Introduction

The 8th International Cybersecurity Data Mining Competition (CDMC 2017) is a challenging, multi-month research and practice competition, focusing on application of knowledge discovery techniques to solve advanced, real-world problems. The competition is associated with the 10th International Workshop on Data Mining and Cybersecurity (DMC2017), which is an associated event to the 24nd International Conference on Neural Information Processing (ICONIP2017), Guangzhou, China.

In this competition, participants are required to solve all of the following tasks, Task 1: Android Malware Classication based on API information, Task 2: Incident Detection over Unied Threat Management (UTM) operation on UniteCloud, and Task 3: Fraud Detection in Financial Transactions. The following sections describe our solution in detail.

2　Task 1: Android Malware Classication based on API information

2.1　Task Description

Software vulnerabilities such as viruses, malware, and other attacks have serious security implications. Android Malware classification is needed to protect our device because of the rapid growth of malware threats for android platform [1].

To install software on the Android operating system, application package (APK) les are used, which includes API (Application Program Interface) information.

The objective of this task is to design a classifier for malware detection based on the API information. A list of APIs obtained by reverse engineering the APK files were provided for the task. The APK files were collected from the Opera Mobile Store [2] over the period of January to September of 2014. The class label of the APK file was determined by the detection results of security appliances hosted by VirusTotal [3]. Adware was not counted as malware in the setting.

The information of the dataset is summarized in Table 1.

2.2 Our Proposed Method

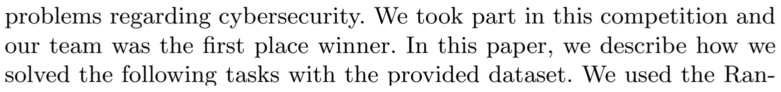
Table 2 shows the example of the API names. The dataset for the task has 37,107 features (APIs) and two labels(1 and -1). 1 stands for a malware and -1 for a benign file. For preprocessing, we separated the API names by a dot character. For example, 'android.accounts.abstractaccountauthenticator.init' returns 'android', 'accounts', 'abstractaccountauthenticator' and 'init'. We only used the last one ('init') as the feature. Some of them are the same name. After preprocessing, we have 10,058 features in total.

We used the Random Forest (RF) classifier with Python scikit-learn [4]. We tuned its hyper parameters to enhance the accuracy of the model. We have selected the best set of hyper parameters for RF. We trained our model data with the hyper parameters. Table 3 shows the list of hyper parameters we used in our experiments.

2.3 Experimental Results

We used 10 fold cross validation to compute the accuracy. In our experiments, we got the accuracy

Table 1. The information of the dataset



problems regarding cybersecurity. We took part in this competition and our team was the first place winner. In this paper, we describe how we solved the following tasks with the provided dataset. We used the Ran-

Table 2. The example of the APIs

18	android.accounts.abstractaccountauthenticator.init
19	android.accounts.account.describecontents
20	android.accounts.account.equals
21	android.accounts.account.hashcode
22	android.accounts.account.init
23	android.accounts.account.tostring

Table 3. Hyper Parameters

hyper parameter	value	hyper parameter	value
bootstrap	True	min_samples_leaf	1
class_weight	None	min_samples_split	2
criterion	gini	min_weight_fraction_leaf	0.0
max_depth	None	n_estimators	100
max_features	1000	n_jobs	1
max_leaf_nodes	None	oob_score	True
min_impurity_decrease	0.0	random_state	0
min_impurity_split	None	verbose	0
warm_start	False		

Table 4. Confusion matrix of a validation dataset

	-1	1
-1	2375	76
1	113	526

Table 5. Confusion matrix of the whole dataset

	-1	1
-1	24106	446
1	473	5872

of 0. 938 with a validation dataset. Table 4 is a confusion matrix of a validation dataset.

We also calculated the accuracy of the whole training data with the best model trained the parameters listed in Table 3. We got the accuracy of 0.970.

Table 5 shows a confusion matrix of the whole training data.

3　Task 2: Incident Detection over Unied Threat Management (UTM) operation on UniteCloud

3.1　Task Description

The incident detection is important for Cloud environments since potential attacks and platform vulnerabilities can pose serious security threats to computers and networks. The objective of this task is to identify various incident accurately from the sensor log les captured from real-time running Unied Threat Management (UTM) on the UniteCloud server [5]. The information of nine selected sensors under the UTM platform was provided [6]. The class label of the log les was determined by incident status determination over the collected log data.

The information of the dataset is summarized in Table 6.

3.2　Our Proposed Method

Table 7 shows the example of the data. The dataset for the task has nine features and two labels (pass, block). For preprocessing, we lled NaN values with 0. V1, V2, V3, V4, V5 and V6 are categorical variables, and the others are continuous.

We excluded V1, V4 and V5.We converted each categorical variable into dummy variables. Table 8 shows the example of the converted V3. The 'gOqV' feature was also excluded because it is not appeared in the test data. After preprocessing, we have 15 features in total. We used the RF classier with Python scikit-learn.

We tuned its hyper parameters to enhance the accuracy of the model. We have selected the best set of hyper parameters for RF. We trained our

Table 6. The information of the dataset

No. of Sample	No. of Features	No. of Classes	No. of Training	No. of Testing
100,000	9	2	70,000	30,000

Table 7. The example of the training data

	V1	V2	V3	V4	V5	V6	V7	V8	V9	label
1	9PsSq	kW	2Cd	mNIpM	IZ	dmOS	62	61	41	pass
2	0wbaV	kW	2Cd	8MXxg	IZ	dmOS	62	72	52	pass
3	J	kW	OP42	5G	EBM	dmOS	46	84	NaN	block
4	xLWCq	kW	2Cd	ZrWjo	IZ	0tBa	63	67	47	pass
5	J	kW	OP42	5G	scP	0tBa	42	84	NaN	pass

Table 8. The converted data (V3)

	04v	2Cd	AtQK	OP42	gOqV
1	0	1	0	0	0
2	0	1	0	0	0
3	0	0	0	1	0
4	0	1	0	0	0
5	0	0	0	1	0

Table 9. Hyper Parameters

hyper parameter	value	hyper parameter	value
bootstrap	True	min_samples_leaf	1
class_weight	None	min_samples_split	2
criterion	gini	min_weight_fraction_leaf	0.0
max_depth	50	n_estimators	100
max_features	auto	n_jobs	1
max_leaf_nodes	None	oob_score	True
min_impurity_decrease	0.0	random_state	0
min_impurity_split	None	verbose	0
warm_start	False		

model with the hyper parameters. Table 9 shows the list of hyper parameters we used in our experiments.

3.3 Experimental Results

We used 10 fold cross validation to compute the accuracy. In our experiments, we got the accuracy of 0.999 with a validation dataset. Table 10 is a confusion matrix of a validation dataset. We also calculated the accuracy of the whole training data with the best model trained the parameters listed in Table 9.

We got the accuracy of 0.999. Table 11 shows a confusion matrix of the whole training data.

Table 10. Confusion matrix of a validation dataset

	block	pass
block	1903	1
pass	0	5096

Table 11. Confusion matrix of the whole dataset

	block	pass
block	18761	6
pass	6	51227

4 Task 3: Fraud Detection in Financial Transactions

4.1 Task Description

Financial fraud is a long standing issue with broad reaching consequences. The goal of this task is to design a classifier for fraud detection based on the financial transaction. The original anonymized data was provided by the financial institution [7], and was synthesized with highly correlated rule based uniformly distributed synthetic data (HCRUD) technique.The transactions from various account and transaction types were provided with 12 features for each transaction. The information of the dataset is summarized in Table 12.

4.2 Our Proposed Method

Table 13 shows the example of the data. The dataset for the task has 12 features and three labels (Non, Fraud, Anon). For preprocessing, we filled NaN values with 0, and 0+5i with 1. V1, V3, V10, V11 and V12 are categorical variables, and the others are continuous. We converted each categorical variable into dummy variables. After preprocessing, we have 32 features in total. We used the RF classifier with Python scikit-learn. We tuned its hyper parameters to enhance the accuracy of the model. We have selected the best set of hyper parameters for RF. We trained our model with the hyper parameters. Table 14 shows the list of hyper parameters we used in our experiments.

4.3 Experimental Results

We used 10 fold cross validation to compute the accuracy. In our experiments, we got the accuracy of 0.978 with a validation dataset. Table 15 is a confusion matrix of a validation dataset. We also calculated the accuracy of the whole training data with the best model trained the parameters listed in Table 14.

We got the accuracy of 0.999. Table 16 shows a confusion matrix of the whole training data.

Table 12. The information of the dataset

No. of transactions	No. of Features	No. of Classes	No. of Training	No. of Testing
100,000	12	3	70,000	30,000

Table 13. The example of the training data

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	label
0	FT	4298	Personal	0	0	0	7	2	1	PM	NaN	Other	Non
1	PA	5070	Home_Loan	5070	0	0	5	5	1	AM	4	AU	Fraud
2	FT	321	Credit	0	0	0	1	2	2	PM	4	Other	Non
3	PA	6488	Personal	0	0	0	3	1	1	AM	NaN	AU	Fraud
4	OTT	9122	Business	0	0	0	2	2	1	PM	4	Other	Non

Table 14. Hyper Parameters

hyper parameter	value	hyper parameter	value
bootstrap	True	min_samples_leaf	1
class_weight	None	min_samples_split	2
criterion	gini	min_weight_fraction_leaf	0.0
max_depth	50	n_estimators	100
max_features	20	n_jobs	1
max_leaf_nodes	None	oob_score	True
min_impurity_decrease	0.0	random_state	0
min_impurity_split	None	verbose	0
warm_start	False		

Table 15. Confusion matrix of a validation dataset

	Anon	Fraud	Non
Anon	52	1	5
Fraud	3	229	2
Non	2	2	404

Table 16. Confusion matrix of the whole dataset

	Anon	Fraud	NoN
Anon	6849	1	5
Fraud	4	21556	2
NoN	3	2	41578

5 Conclusion

We took part in the CDMC2017 competition, and our team got the first place winner. For all the tasks, we adopted the Random Forest classifier commonly used in machine learning. Our hyper-parameter tuning and feature selection enhanced classification accuracy, which is high enough for real-world problems.

Acknowledgments.

We thank Marie Ohki, Miku Kabeyama, Nagisa Kawai, Juri Koumoto, Midori Sakai, Haruka Nakai, Mayu Nagao, Mayu Nishimoto, Ayumi Hirao, Ririko Hirao and Jun Yamauchi for their contributions to this work.

References

1. Tao Ban, Takeshi Takahashi, Shanqing Guo, Daisuke Inoue, Koji Nakao. Integration of Multi-modal Features for Android Malware Detection Using Linear SVM. The 11th Asia Joint Conference on Information Security (ASIAJCIS 2016), Fukuoka, Japan, Aug. 2016.
2. Opera Mobile Store. Available: http://html5.oms.apps.opera.com/. [Accessed: 29-Aug- 2017]
3. VirusTotal. Available: https://www.virustotal.com/en/. [Accessed: 29- Aug- 2017]
4. Pedregosa et al. Scikit-learn: Machine Learning in Python. JMLR 12, pp. 2825-2830, 2011.
5. UniteCloud. Available: http://www.unitecloud.org/. [Accessed: 29- Aug- 2017]
6. Shaoning Pang, Tony Shi, Ruibin Zhang and Denis Lavrov. 2017 CDMC Task 2: Incident Detection over Unified Threat Management (UTM) operation on Unite-Cloud. Unitec Institute of Technology, Auckland, New Zealand, 2017.
7. Internet Commerce Security Laboratory (ICSL). 2017 CDMC Task 3: Fraud Detection in Financial Transactions. Federation University Australia, Ballarat, VIC, Australia, 2017.