International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

Improving Malware Detection Classification Accuracy with Feature Selection Methods and Ensemble-based Machine Learning Methods

P HarshaLatha, R Mohanasundaram

Abstract*: *Malware is evolving serious threats to internet security. The classification of malware is extremely crucial in recent days. The traditional models are failed to achieve to get effective accuracy rate and the machine learning models are the basic models that accomplish the task of classification in a certain way, but in recent decades malware attacks are very drastic and difficult to achieve zero-day attacks. To compete with new malware, ensemble methods are highly effective and give better results of accuracy. In this paper, we propose a framework that combines the exploit of both feature selection methods and ensemble learning classifiers and gives better results of classification. In the experimental results, we prove that this combination of methods gives better classification with high accuracy of 100% with the Random Forest ensemble classifier.

Keywords* : *Machine Learning, Feature Selection methods, Classification, Malware detection, Ensemble Learning

I. INTRODUCTION

Malware is malicious software that steals information from the user computer and may lead to personal loss or damage to information. According to AV-TEST statistics [1], consistently, this Institute enrolls more than 350,000 new malware and Potentially Unwanted Applications (PUA). On the latest update of total malware by AV-TEST are 953.27 millions of malicious ones.

There are different types of malware like virus, adware, spyware, rootkits, worms, bugs, bots, Trojan horses, ransomware etc. [2]. To study this malware, two types of analysis are there. The first one is static one and second one is dynamic analysis. The static analysis deals with malicious code without executing it. The dynamic analysis deals with malicious software while executing it [3]. In the same way, identifying the new malware whether it is malware or not, generally, it is needed for classification. The classification of malware is a crucial one to identify new malware. There are several advancements takes place in the classification after arrival of feature selection and machine learning or data mining techniques.

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author
P. HarshaLatha*, Research Scholar, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamilnadu, India. Email: harsha17latha@gmail.com
Dr. R. Mohanasundaram, Associate Professor, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamilnadu, India. Email: mohanasundaramr@vit.ac.in

The feature selection method plays a vital role in classification and it saves the processing time of classification, it improves the accuracy, it reduces the dimensionality of data and removes irrelevant features for classification [4].

Machine learning in malware detection classification brings more advancement in classification [5]. Especially ensemble learning improves the results of machine learning in a better way. Ensemble learning gives optimal solutions by combining base models of machine learning. Section 3 gives an overview of ensemble learning.

In this paper, we propose an approach that improves the accuracy of the classification. Some of the researchers prove that feature selection techniques improve the classification accuracy and some researchers prove that ensemble learning techniques increase the accuracy performance in classification. In this study, we present an approach that the combination of these two mechanisms feature selection and ensemble learning gives better performance and best results of classification, especially in the case of accuracy.

The related literature survey is presented in section 2. In section 3, an overview of the feature selection and ensemble classifiers are discussed. The proposed work of the paper is presented in section 4. The experimental results are discussed in section 5 and section 6 is conclusion of this work.

II. RELATED WORK

In the comparative study [6] about six different types of feature selection methods with four different machine learning classifiers Neural networks, Support Vector Machine (SVM), J48 and Naïve Bayes give the best result of 97% accuracy with the combination of PCA feature selection with SVM.

[7] presents a novel approach for malware detecting classification with machine learning. Here Chi-square and RF methods are used as feature selection methods. The machine learning classifiers are used to classify the features are K-Nearest Neighbor (KNN), Decision Tree (DT), SVM and RF. Among them, DT gives the best accuracy of 99.11%. The classification of malware families is performed by this novel approach. [8] Surveyed the research papers of machine learning techniques on malware detection. The paper gives an overview of all the feature selection methods and machine learning algorithms. Among all the methods and techniques of feature selection, it says that Random Forest provides better results of accuracy and SVM gives better classification.

the previous research works, it is noticed that the feature selection methods and machine learning techniques give better results in malware detection classification.

In this paper, we enhance the work that takes the combination of feature selection methods and ensemble learning to prove better results in classification where the results

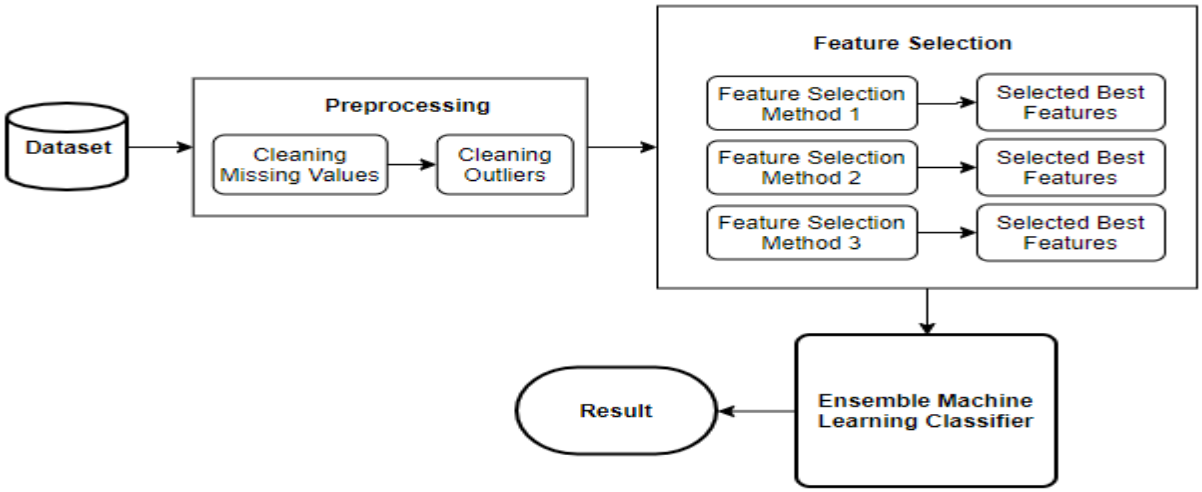

Fig. 1.Proposed Model

ensemble learning is the basic combination models of machine learning techniques. Hence, this paper gives better with ensemble methods which are enhancement of machine learning along with feature learning methods.

III. OVERVIEW OF METHODS

Three different feature selection methods and four different ensemble learning methods are used in this proposed work. The overview of those methods are given below.

A. Feature Selection Methods

Feature selection methods are widely useful for many purposes like dimensionality reduction, improve accuracy, remove irrelevant features and reduce the computational or processing time [4,9,10]. In this paper, three feature selection methods are used which are ExtraTreesClassifier, Percentile and KBest methods.

ExtraTreesClassifier is a tree-based ensemble classifier. It is very similar behavior of Random Forest ensemble learning. It combines the group of decision trees known as "forest" and produce the classification output [11]. Each decision tree is formed with the training sample. Each tree produces some k number of features from the training sample. The decision is taken by calculating the formulas of Information Gain and Entropy. The (1) and (2) represents the formulas for Information Gain and Entropy respectively [12].

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|Sv|}{S} Entropy(Sv) \quad (1)$$

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2(pi) \quad (2)$$

Where,
S : a training set
p_i : the proportion of rows with output label is i
c : Number of Unique Class labels

The **percentile** [13] method is a feature selection method that selects the relevant features according to the highest percentile scores of the features.

KBest [14] method is a feature selection where K is the number of best features and it selects features basing upon the highest scores of feature importance scores.

B. Ensemble Learning Methods

Ensemble learning is a combination of several basic models of machine learning. It gives the optimal solution to the problem [15, 16]. In this paper, four different ensemble learning methods are used which are Random Forest [17, 18], Bagging, Gradient Boosting and AdaBoost [18].

IV. PROPOSED WORK

Fig. 1 gives an overview of the proposed work of this paper. In this proposed work, steps involve like pre-processing of dataset, selection of relevant features by using feature selection methods and applying Ensemble learning classifier to produce better results. In the pre-processing step, missing values and outliers are detected and removed using the WEKA tool. Second, Feature selection methods which reduces the dimensionality and provides solution to over-fitting problem and also it reduces the dimensionality feature space which produces the most relevant features for classification. Third, deals with the machine learning classifier. The classifier is an ensemble learning model. Finally the combination of feature selection methods and ensemble learning methods produces better accuracy and less processing time. The Python script is used for feature selection and classification.

The step by step process of Pseudo-code for proposed model is given below.

1. Select Dataset D

2. Preprocess the dataset and remove missing values and outliers from the dataset D
3. Impute the feature selection method ExtraTreesClassifier on dataset D
4. Extract the best features from dataset D basing on the feature importance and form new dataset D1
5. Build a ensemble machine learning model with classifiers Bagging, Gradient Boosting, Random Forest and AdaBoost
6. Compute Results
7. Repeat from step 3 to step 6 for feature selection methods Percentile and KBest instead of ExtraTreesClassifier

Dataset Description

The information about the dataset is collected from an Internet source Kaggle. This dataset contains a total 100000 records of malware samples. Among them 50000 samples are benign and 50000 samples are malware. It contains 35 features which are used for feature selection and classification.

Feature Selection

For all the features in the dataset, three different feature selection methods ExtraTreesClassifier, Percentile, and KBest methods are applied respectively and select the best features among all 35 features. ExtraTreesClassifier selects 13 best features among 35 features. Percentile selects 4 best features among 35 features and KBest selects 10 best features from the total 35 features. Fig. 2. gives an overview of feature selection methods.

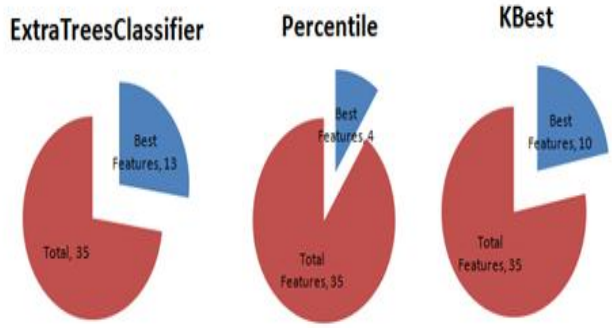

Fig. 2.Number of Features selection by Feature Selection Methods

The best features are selected by feature importance by the particular method respectively. Fig. 3, 4 and 5 gives the feature importance of the selected best features by the particular feature selection method respectively.

The selected best features are used for classification with ensemble learning classifiers.

Classification

The classification is performed by using four different ensemble machine learning classifiers that are Bagging, AdaBoost, Gradient Boosting, and Random Forest. The below Table. describes the accuracy of ensemble classifiers along with relevant feature selection methods. The ensemble machine learning always gives the best optimization solution for the problem. Four different ensemble learning classifiers are used to test the performance of the model those are Adaboost, Bagging, Gradient Boosting and Random Forest.

All four classifiers are applied to all the three feature selection methods and give better accuracy results. Table1 shows the Accuracy of all classifiers along with all three feature selection methods.

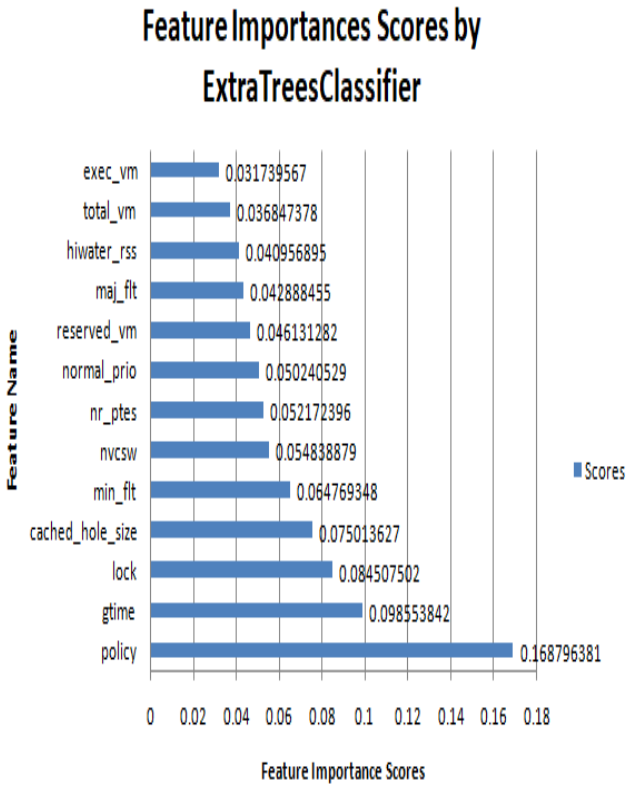

Fig. 3.Feature Importances Scores by ExtraTreesClassifier.

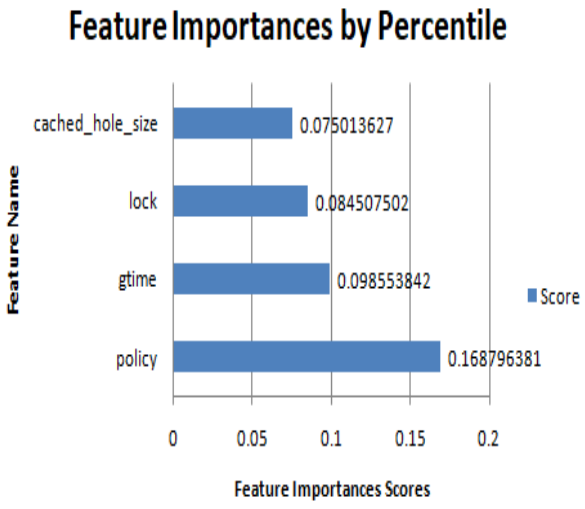

Fig. 4.Feature Importances Scores by Percentile.

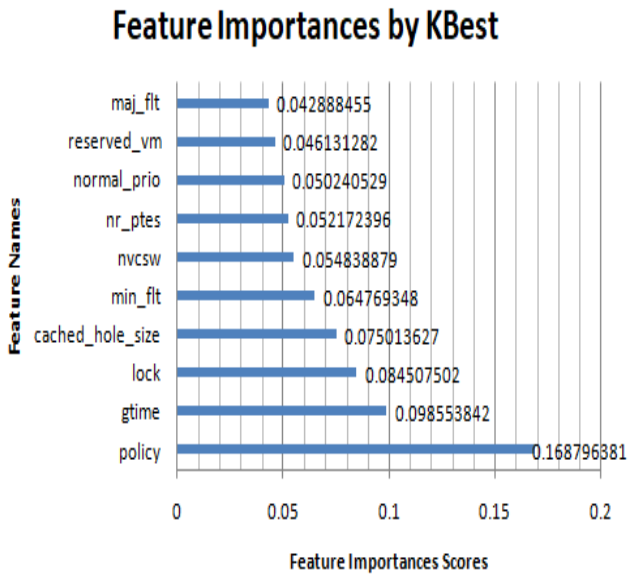

Fig. 5.Feature Importances Scores by KBest.

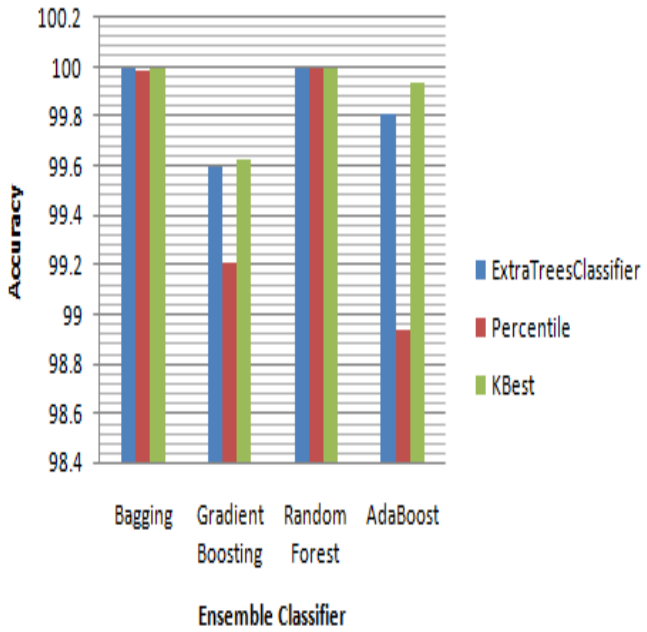

Fig. 6.Accuracy Chart.

Among all four classifiers, the Random forest gives 100% accuracy in all three states of feature selection methods. The chart in figure demystifies the performance of the model with accuracy basing along with all four classifiers and with all three feature selection methods.

Table- I: Accuracy (in Percentage) of Ensemble classifiers with feature selection Methods.

S.No.	Ensemble Classifier	Extra Trees Classifier	Percentile	Kbest
1	Bagging	100%	99.99%	100%
2	Gradient Boosting	99.60%	99.21%	99.63%
3	Random Forest	100%	100%	100%
4	AdaBoost	99.81%	98.94%	99.94%

V. RESULT AND DISCUSSION

Feature selection methods are used to select the best relevant features for classification. In this process it produces the set of best features. The ensemble learning classifiers are used for classification. One of the performance metrics of classification is Accuracy; it is calculated from the confusion matrix. Figure 6 illustrates about accuracy of classification with all three feature selection methods. The experimental result of this research gives better accuracy in the combination of feature selection methods and ensemble classifier of Random Forest which gives 100% accuracy. Table 1 gives result of the methods in accuracy metric with all combinations of the methods of feature selection and ensemble classifiers respectively.

VI. CONCLUSION

The previous existing literature is not that much efficient of getting good accuracy in classification. To compete with recent malwares especially in classification there is advancement in applying methods. In this paper, we used advance methods of feature selection and ensemble methods for classifying malware. We present framework that shows the importance and benefits of using both feature selection methods and ensemble learning classifiers in classification of malware. The experimental results give high accuracy of 100% using ensemble learning classifier Random Forest. All three feature selection methods are given good results and also by ensemble learning classifiers. The processing time for classification is also reduced because of removing irrelevant features in feature selection process for classification.

REFERENCES

1. AV-TEST (2019, November 29). The Independent IT-Security Institute [Online]. Available: https://www.av-test.org/en/statistics/malware/
2. Ray, A., & Nath, A. (2016). Introduction to Malware and Malware Analysis: A brief overview. International Journal, 4(10).
3. Gandotra, E., Bansal, D., & Sofat, S. (2014). Malware analysis and classification: A survey. Journal of Information Security, 5(02), 56.
4. Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16-28.
5. Ucci, D., Aniello, L., & Baldoni, R. (2017). Survey on the usage of machine learning techniques for malware analysis. arXiv preprint arXiv:1710.08189.
6. Khammas, B. M., Monemi, A., Bassi, J. S., Ismail, I., Nor, S. M., & Marsono, M. N. (2015). Feature selection and machine learning classification for malware detection. Jurnal Teknologi, 77(1).
7. Sethi, K., Kumar, R., Sethi, L., Bera, P., & Patra, P. K. (2019, June). A Novel Machine Learning Based Malware Detection and Classification Framework. In 2019 International Conference on Cyber Security and Protection of Digital Services (Cyber Security) (pp. 1-4). IEEE.
8. El Merabet, H., & Hajraoui, A. (2019). A Survey of Malware Detection Techniques based on Machine Learning. INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS, 10(1), 366-373.

9. Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2018). Feature selection: A data perspective. ACM Computing Surveys (CSUR), 50(6), 94.
10. Mays, M., Drabinsky, N., & Brandle, S. (2017). Feature Selection for Malware Classification. In MAICS (pp. 165-170).
11. Scikit-learn (2019, November 29) [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html
12. Geeks for Geeks. (2019, November 29). A Computer Science Portal for Geeks [Online]. Available: https://www.geeksforgeeks.org/ml-extra-tree-classifier-for-feature-selection/
13. Scikit-learn. (2019, November 29). [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectPercentile.html
14. Scikit-learn. (2019, November 29). [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html
15. Idrees, F., Rajarajan, M., Conti, M., Chen, T. M., & Rahulamathavan, Y. (2017). PIndroid: A novel Android malware detection system using ensemble learning methods. Computers & Security, 68, 36-46.
16. Feng, P., Ma, J., Sun, C., Xu, X., & Ma, Y. (2018). A Novel Dynamic Android Malware Detection System With Ensemble Learning. IEEE Access, 6, 30996-31011.
17. Alam, M. S., & Vuong, S. T. (2013, August). Random forest classification for detecting android malware. In 2013 IEEE international conference on green computing and communications and IEEE Internet of Things and IEEE cyber, physical and social computing (pp. 663-669). IEEE.
18. Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1249.

AUTHORS PROFILE

P HarshaLatha received her Bachelor's Degree in Computer Science from Sri Padmavathi Degree College, Tirupati, Andhra Pradesh, India in 2008. She received Post Gradute degree in Master of Computer Applications from Sreenivasa Institute of Technology and Management Studies, Chittoor, Andhra Pradesh, India in 2011. She has working experience as a Software Engineer in MNC Company from 2011 to 2013. She is currently pursuing M.Tech by Research as a Full time Research Scholar in the School of Computer Science and Engineering at Vellore Institure of Technology, Vellore, Tamilnadu, India. Her areas of interest are Machine Learning, Malware Analysis, Big Data, Cyber Security, Artificial Intelligence, Information Security.

Dr. Mohanasundaram Ranganathan received his B.E. degree in Electrical & Electronics Engineering from Velalar College of Engineering and Technology, Erode in 2006, M.E. Embedded System Technologies from Velalar College of Engineering and Technology, Erode, Tamilnadu, India in 2008 and Ph.D. from Anna University, Chennai, India in 2015. He is currently as Associate Professor in the School of Computing Science and Engineering at Vellore Institute of Technology (Deemed University), Vellore, Tamilnadu, India. He has published more than 35 research papers in International and National Journals/Conferences. He have contributed few book chapters in the area of advanced embedded systems and swarm intelligence. His areas of interest are Swarm Intelligence, Wireless Sensor Networks, Computer Networks, VLSI, Embedded Systems, Mobile Communications, Life science.