A STUDY ON THE INTELLIGENT METHOD FOR DETECTION OF COMPUTER VIRUSES

*Abdumuminov Abdurafiq Abdurashidovich
**Ibragimov Jalaliddin Obidjon o'g'li
*** Shoraimov Khusanboy Uktamboyevich
*Republican center for management of telecommunications networks of Uzbekistan. SUE.
** Teacher of the Department, “Systematic and Practical Programming”, Tashkent University of Information Technologies named after Muhammad Al-Khwarizmi, UZBEKISTAN
*** Teacher of the Department, “Systematic and Practical Programming”, Tashkent University of Information Technologies named after Muhammad Al-Khwarizmi, UZBEKISTAN

Abstract—This paper makes a virus detection study based on the D-S theory of evidence, which applies to two types of classifiers, support vector machines and probabilistic neural networks to detect the virus. Then, the D-S theory of evidence is used to combine the contribution of each individual classifier to obtain the final decision. The experiment tests and result analyses demonstrate that it is efficient for unknown viruses and variant viruses to improve accuracy rate of integration virus detector by using D-S theory to create the isomeric classifier.

Keywords- *computer viruses; virus detection; D-S theory of evidence; classifier; credit distribution*

I. INTRODUCTION

The daily overflowing computer virus is one of the most serious menaces for information security. It is not efficient for traditionally characteristic scan methods because of encrypted viruses and anamorphic viruses so that the studying new anti-virus method is very urgent. The paper presents a new method of combining the dynamic detection of computer viruses with static detection of computer viruses based on D-S theory of evidence to much improve accuracy rate of computer virus detection.

The results of distinct classifiers are combined, but selection the best classifier, to improve the performances. It is beneficial for the combination of classifiers since the complementary information of different classifiers may be given[1]. Many scholars have widely studied the combination of support vector machines to certify that it is much better than single support vector machine and to obtain satisfactory results.

The multiple support vector machines are used as the member classifier to model for the dynamic behaviors of the viruses in the virus detection system. And, the multiple probabilistic neural networks are used as the member classifier to model for the static behaviors of the viruses. Then, the test results of various member classifiers are combined by D-S theory of evidence to form the final test conclusions.

II. Virus Test Engine Based on D-S Theory of Evidence

A. System Framework

The virus detection system framework based on D-s theory of evidence is presented on figure 1. The dynamic behavior features and static features are systematically comprehensively considered. The two kinds of eigenvectors are extracted to demonstrate the model of sample programs. One is API functions cited by programs. Another is n-gram information statically extracted by PE programs. It can detect and analyze the behaviors of programs to effectively detect unknown viruses and all kinds of polymorphic viruses. The probability statistics method is applied to mining implicit information from the n-gram set in the static analysis process to detect automatic production machine, compiler, programming environment and even some of the programming author habits of programming viruses and to effectively prevent virus author counterattack. There are Probabilistic Neural Network member classifier (PNN) and support Vector Machine members classifier (sVM) in the system. The combination between PNN and sVM is based on D-s theory of evidence.

The characteristics of PNN are used as the called information of API function and the characteristics of sVM are n-gram information of the program in the training member classifier process to enlarge the difference and correlation of the member classifiers.

B. Realization Method

There are Boosting and Bagging in generating methods of usually used member classifiers. Bagging is based on repeatable sampling some examples randomly from the original training sets to train the individual member classifier. It has increased the difference of the individual classifier by repeating selection training sets to improve its generalization ability. The training sets of each individual member classifier are decided by the generated classifier performances in the Boosting method so that the examples wrong judged by classifiers will great probabilistically show up in new classifiers. The Bagging method is selected to generate the individual member classifier for detection system time consuming because the trainings of member classifier generated by the method may be parallel disposed. It is as following:

The training set is given and the series of training subsets S1, S2, ... and ST are obtained by repeated sampling. Then, the information gain algorithm is used to select the static attributes for playing an important role of classification as PNN input to train the individual members of PNN on all kinds of training subsets and to pick out the dynamicproperties for playing an important role of classification as SVM input to train the individual members of SVM classifiers.

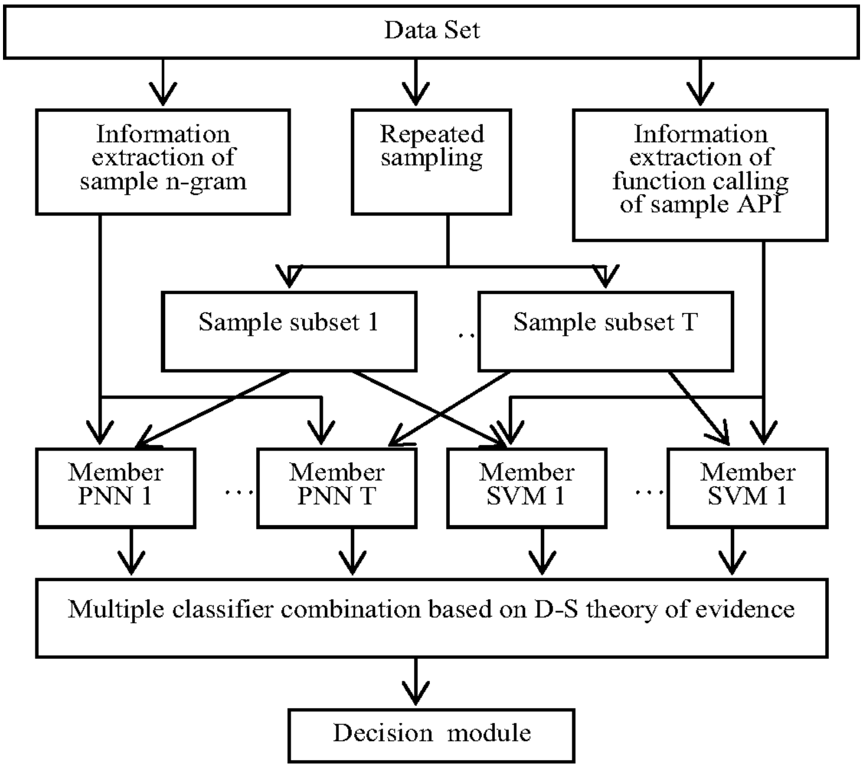
Data Set
Information extraction of sample n-gram
Repeated sampling
Information extraction of function calling of sample API
Sample subset 1
..
Sample subset T
Member PNN 1
...
Member PNN T
Member SVM 1
...
Member SVM 1
Multiple classifier combination based on D-S theory of evidence
Decision module

Figure 1. Virus detection system framework based on D-S theory of evidence

B. Credit Distribution Method Based on Distance Measure between Classifier

The credit distribution method discussed above is based on the classification performances of the classifiers. However, it is much related to selection of test sets so that it is not very stable. The distance between classifiers should be enlarged for all kinds of classifiers during actual modeling whatever its working principles. It means that the more detachable for the classifiers, the better results of the classification so that the distances between classifiers are chosen as the evidence credit distribution basis of all member classifiers.

N classifiers $e^{(1)}$, $e^{(2)}$,...$e^{(N)}$ are applied to K classifications. Each classification is expressed as 9_k , k=1,2,...,k. And, the recognition framework is $9=\{9_1, 9_2, .9_k\}$ under the D-S theory of evidence.

Suppose that x_k is the training sample matrix of 9_k classification, K=1, 2, ..., K. The characteristic matrix of extracted from feature selection modules of different classifiers is marked as $X_k^{(n)}$. All kinds of classifiers are isomorphism or heterogeneous and their character spaces may be different.

The sample expression is abstracted as a modeling function formula (5) for each classifier under diverse character spaces.

$$\mathbf{r^{(n)}(XkHk^{(n)}\ ,\ K=1,\ 2,\ \ldots\ ,\ K,\ n=1,\ 2,\ \ldots\ ,\ N\ (5)}$$

Of course, all kinds of modeling expression are different according to diverse classifier Principles and methods.

The modeling of each member classifier for test sample x is expressed as formula (6).

$$\mathbf{r^{(n)}(X)=I^{(n)}\ ,\ n=1,\ 2,\ \ldots\ ,\ N \qquad (6)}$$

According to different expressions between training samples and test samples, the distance between training samples and test samples is calculated by formula (7) for each classifier $e^{(n)}$.

IV Experiment Results and Analysis

There are 845 samples in data set. The 509 samples are normal programs and the 336 samples are virus programs. The n-gram information is extracted for all samples to select characteristics.

There are SVM member classifiers and PNN member classifiers in the integrated classifiers. The virus static detection method, the dynamic testing method and both combined test method are compared in experiments. The results are demonstrated on Table 1.

The conclusion is gained from experiment results as following:

The combination method between dynamic virus detection and static virus detection is much more accurate than one of them because the characteristics of the integrated classifiers are API information and n-gram information of programs. Both of them are no correlation so that the difference between various member classifiers is maximum expanded and the performance of detection system is much improved.

V. Conclusions

The combination between the dynamic virus detection technology and the static virus detection technology is presented here after there are the thought advantages and disadvantages respectively for each of them. The results of various detectors are combined by D-S theory of evidence because different member classifiers are isomeric in the system so that they are not suitable for the traditional ballot combination. The different types of eigenvectors are applied to the training member classifiers on the combined detectors to increase the none-dependency and difference of the member classifiers so that the detection precision rate of the integrated classifier is much improved.

References

1. Devijver P A, Kittler J. Pattern Recognition, A Statistical Approach. London: Prentice Hall,2020.
2. Dimitrios S, frossyniotis, Andress S. A Multi-SVM Classification System, Proceedings of the Second International Workshop on Multiple Classifier Systems (MCS 2001), LNCS, Vol 2016, 2001, 198-207.
3. Kim H, Pang S, Je H, et al. Construcing Support Vector Machine Ensemble. Pattern Recognition, 2018, 36(12): 2757-2767.
4. Breiman L. Bagging Predictors. Machine Learning, 1996, 24(2):123- 140.
5. Schapire R E. The Strength of Weak Leanability. Machine Learning, 2013, 5(2): 197-227.
6. Xu L, Krzyzak A, Suen C. Methods of combining multiple classifiers and their applications to handwritten recognition. IEEE Translactions on Systems, Man and cyberneties, SMC, 2012, 22(3):418-435
7. Askarov B., Yuldashev A., Sultanova D. SYNERGY METHOD FOR SOLVING SOME PROBLEMS OF EDUCATION //ASJ. – 2021. – Т. 2. – №. 56. – С. 15-19.
8. Sultanova D. T. PROSPECTS FOR THE DEVELOPMENT OF TOURISM IN UZBEKISTAN //Экономика и социум. – 2021. – №. 3-1. – С. 289-292.
9. Zufarjonovna J. G. USING WEB-QUEST TECHNOLOGY IN ENGLISH LESSONS AS FOREIGN LANGUAGE //INTERNATIONAL JOURNAL OF SOCIAL SCIENCE & INTERDISCIPLINARY RESEARCH ISSN: 2277-3630 Impact factor: 7.429. – 2022. – Т. 11. – С. 161-164.
10. Zufarjonovna J. G. BENEFITS OF USING WEB-QUEST TECHNOLOGY IN ENGLISH LESSONS AS FOREIGN LANGUAGE //INTERNATIONAL JOURNAL OF SOCIAL SCIENCE & INTERDISCIPLINARY RESEARCH ISSN: 2277-3630 Impact factor: 7.429. – 2022. – Т. 11. – С. 158-160.