様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月 ２３ 日現在

機関番号：８２６３６

研究種目：若手研究(B)

研究期間：2012～2014

課題番号：２４７０００８４

研究課題名（和文）配列解析による知能マルウェア対策スキームに関する研究

研究課題名（英文）An intellectual anti-malware scheme using advanced sequence analysis techniques

研究代表者

班　涛（Ban, Tao）

独立行政法人情報通信研究機構・ネットワークセキュリティ研究所・サイバーセキュリティ研究室・主任研究員

研究者番号：８０４６２８７８

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：様々なステルス技術を備えたマルウェアの検出率を高める為、本研究では、配列解析に基づいた計算知能技術をマルウェアから生成した配列に適用し、信頼性の高い検出・分類手法について研究開発を行った。新たに提案した編集距離カーネル関数とスペクトラムカーネル関数では、配列の類似度を定量的に評価することが可能になり、それらのカーネル関数を最先端のサポートベクトルマシン分類器に組み込むことによって、高精度且つ効率的にマルウェアを検出できる分類器の実現が可能になる。本提案手法は、２５種類のパッカーによる難読化された検体データベースで実証実験を行い、99%以上の精度でパッカー特定ができることを実証した。

研究成果の概要（英文）：Research and development on computational intelligence techniques based on advanced sequence analysis are pursued in the aim of an analysis system that can detect polymorphic malware programs with good accuracy and efficiency. The newly proposed edit distance kernel function and spectrum kernel function make it possible to quantitatively evaluate the degree of similarity between sequences. Incorporating these kernel functions to the state-of-the-art classifiers, such as the support vector machine, renders the creation of a practical malware detection system possible. The proposed methods are evaluated using a database comprised of obfuscated programs generated by 25 types of packers. Their effectiveness and efficiency are illustrated by prediction accuracies over 99% and very quick system response time.

研究分野：サイバーセキュリティ

キーワード：マルウェア解析　配列解析　パッカー特定　サポートベクトルマシン　カーネル関数　スパムメール

１．研究開始当初の背景

Recent research on malware analysis indicates that the difficulties in fighting against evolving polymorphic malware programs could be addressed by modern Computational Intelligence (CI) methods. The key to the success of CI methods lies in the effective exploitation of ordering information within binary codes (in static analysis perspective) or runtime behaviors (in dynamic analysis perspective) of the programs. According to our feasibility studies, incorporating sequence-alignment kernels to advanced learning algorithms such as support vector machine (SVM) may result in the-state-of-the-art performance in many related tasks.

Despite of the significant performance improvement obtained by using sequence-alignment-kernel based classification algorithms, our feasibility study also revealed the computational intensive nature as well as the difficulties to capture the information from common substrings of adaptive length for these kernels. This motivated the research on an efficient substring kernel function that supports fast similarity evaluation and dynamic substring length adjustment.

２．研究の目的

In the aim of a practical solution to growing cyber threats induced by the polymorphic malware, we conducted research on incorporating advanced sequence-alignment techniques in kernel-based learning. This study comprises the following elements.

(1) Design and implementation of an improved n-gram kernel function that supports dynamic n value, high computational efficiency, and other advanced features such as adaptive weighting.

(2) Evaluations of proposed anti-malware scheme base on the improved n-gram kernel function and other available kernel functions.

(3) Extension of the proposed scheme to related security problems such as spam email filtering that are approachable using this scheme.

３．研究の方法

By implementing the maximal margin principle, Support vector machine (SVM) has demonstrated the-state-of-the-art generalization performance on many real world applications. A key feature that SVM has introduced is called the kernel trick: With a so-called kernel function, i.e., a similarity function over pairs of data points in raw representation, SVM implicitly maps the inputs into a high-dimensional feature space and performs linear classification therein. The kernel trick enables the modularity of a learning task in its dual form: on the one hand, there is no need to change the underlying algorithm to accommodate a particularly chosen kernel function; on the other hand, other types of pattern analysis could also be substituted while retaining the chosen kernel. Typically, the pattern analysis component is of general purpose and the kernel function shall reflect the specific data type and domain knowledge in the data domain under discussion.

The focus of our study is to define kernel functions that could best measure the similarity of pair of malware programs by exploiting the sequential nature of the data and our domain knowledge. We have performed the following research under the scope of this study through fiscal year (FY) 2012 to 2015.

(1) Literature survey on theories, algorithms, and applications of substring/subsequences analysis,

kernel based learning methods, and multiple kernel learning.

(2) Design and implementation of the Dynamic-Length SubString (DLSS) kernel function to satisfy the requirements for ultra-fast classification on large-scale databases.

(3) Incorporation of the kernels in learning models and formulation of a classification system with model creation, parameter tuning, prediction, and performance evaluation functionalities.

(4) Study on malware analysis from static analysis aspect. The DLSS kernel and other kernels are applied to identify the encoding packers, where a binary-code segment storing the packer information will be the input to the classification system.

(5) Application of the proposed scheme to spam email analysis. Study has been performed on the spam email database, where double bounce emails sent to an email server are gathered. Analysis covered two-class classification, i.e., dividing the email corpse into malicious spams and unhalmful spams, and other exploratory analysis using methods such as kernel-based clustering.

４．研究成果

(1) Algorithm of DLSS and its application to malware packer identification

Packers are software wrappers that put around pieces of software to compress and/or encrypt their contents. Malware authors today tend to take advantage of easily acquirable packers to transform their malware programs to evade detection from signature-based anti-virus (AV) scanners. As the most popular obfuscation technique, packing facilitates the composition of new undetectable malware variants by recursively applying a handful of packers to a malware program. This has resulted in an exponential increase in the diversity of malware programs and therefore significantly degrades the detection rate of signature-based AV scanners.

A common way to cope with the packer problem is to extract the original code from the packed program using an appropriate unpacker prior to further analysis. To this end, numerous unpackers have been devised based on reverse engineering the particular packers. Generally, unpacking is done by monitoring the execution process of the program and capturing the memory snapshot at a right timing. As an unguided unpacking attempt will expose the system to potential malware damage, the unpacking operation has to be performed in an isolated sandbox environment that supports critical system resource isolation and prompt system recovery. Given the large pool of available unpackers and the hardships in manipulating their proprietary graphical user interfaces (GUIs), directly testing a malware database of thousands of specimens against all available unpackers will entail significant engineering and computing efforts. To reduce the cost of unnecessary unpacking operations, packer identification – an intermediate step to diagnose the packer that created the given program without physically executing the program – is suggested in recent studies.

In our work (journal paper ③, conference paper ⑨), an SVM based on Levenshtein-distance kernel (LDSVM) was first introduced for packer identification. In LDSVM, the most packer-relevant parts, i.e., the entry point containing sections of packed programs are taken as the input to the classifier. During training, a radio basis function (RBF) induced by the Levenshtein distance upon the code segments is employed as the proximity metric to evaluate the similarity between packers. Then an SVM classifier is built upon the training dataset and evaluated on testing malware specimens. LDSVM is proven to be a promising scheme to bring together the advantage of signature-based and machine learning based approaches, duly supported by its preferable performance in empirical studies.

In another recent work (conference paper ⑧), we proposed to use a kernel function that supports dynamic-length substring (termed p-spectrum kernel in the paper) to replace the LD kernel, and thus the high computing cost of LDSVM can be alleviated. The advantages of adopting a spectrum kernel instead of the LD kernel include: First, it helps to save computation cost in training and testing, leading in improved scalability and usability. The time complexity in training an SVM with spectrum kernel of order p is $\mathrm{O}\,(\mathrm{NL}\,\log L\, +\, \mathrm{N\,L})$, compared with $\mathrm{O}\,(\mathrm{L}^2 N^3)$ for LDSVM. Here, N is the number of samples in the training dataset, and L the length of the input code segments. The time complexity for prediction will be $\mathrm{O}\,(\mathrm{L}\,\log L)$, which is a reasonable cost for real-time responding end-user clients. Second, it enables the exploration of signatures that are located apart from the starting point of the entry point containing section, and therefore could cope with packers with more complicated nature. Finally, treatment of the unique p-spectrum as independent features enables the application of advanced feature selection methods, which can not only enhance the accuracy of the system but also provide valuable hints to discover the most representative signatures of the malware.

Since performance of the algorithms usually depends heavily on the parameters, we use 10-fold cross validation to tune the parameters of the compared algorithms. As an example of the parameter tuning process, Figure 1 shows how the length parameter L is determined for p-spectrum SVM (pSSVM). The dark solid line shows the overall prediction accuracy averaged from the top-10 parameter settings for each L parameter. And the green dashed line shows the averaged training time for the building the model using the same setting. It is easy to observe the increment in accuracy as L increase from 40 to 320. Meanwhile, the training time keeps increasing as L increases to 280, and decreases to some extent for L = 320. These results suggest that 320 is the most appropriate length value: it yields the best generalization performance on a wide range of parameter combinations with a reasonable computing cost: training performed on 3,228 samples with 26 classes takes about 2 seconds.

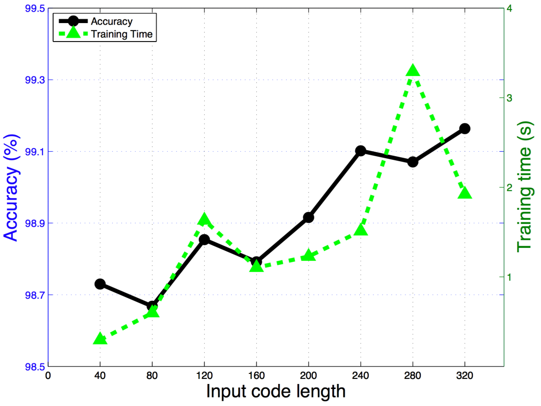

Figure 1. Tuning parameter L using 10-fold cross validation. L is selected from {40, 80, 120, 160, 200, 240, 280, 320}.

In Table I, we compare the overall classification accuracy using the best results returned by 10-fold cross validation. When LD is applied as the similarity metric, kNN yields an overall accuracy of 98.64%. Due to the simple decision rule for kNN, we can attribute the improvement to LD for being able to capture the most essential discriminating information for different packers. By virtue of the good generalization ability of SVM, it is not surprising that the accuracy is further improved to 99.35% using LDSVM.

When the p-spectrum kernel is applied to the learning, we got an overall accuracy of 99.19%, which is comparable with LDSVM's generalization performance. And it shows a 7.5% percent improvement in overall accuracy as compared to SUNKNN – a conventional method – over performing the rest of evaluated methods. This indicates that the discriminant information captured by the p-spectrum kernel is as much as LD. While retaining the accuracy at the same level as LDSVM, it substantially reduces the computational cost to a few seconds, speeding up the training of LDSVM by three orders of magnitude.

Table I. Cross validation results based on parameter tuning

CLASSFIER	ACC.(%)	TRAIN. TIME(S)
SUNKNN	91.42	–
SUNSVM	93.93	30
LDKNN	98.54	–
LDSVM	**99.35**	2136.7
pSSVM	99.19	**1.9**
PEiD (normal)	69.81	–
PEiD (deep)	63.5	–
PEiD (hardcore)	63.5	–

(2) Extended application to spam email analysis

E-mail is an important and efficient means to communicate in today's digital life. However, because of their convenience, spammers frequently abuse emails for commercial, political, and other purposes. Recent spam emails tend to be sent by various malware (e.g., bots, worms). Such kinds of spam emails contain URLs which links to webservers for purpose of diverse cyber attacks, e.g. malware infection, user information theft, phishing attacks, etc. We call such spams malicious spam emails (MSEs) which need to be differentiated from less harmful ones that are merely for advertising purpose.

In our recent work (journal paper ②, conference paper ①⑤⑦), we present a study on online MSE detection using SVM with string kernel functions. Comparison study is performed on popular classification algorithms using generalization performance measures including accuracy, precision, recall, and F1-score. The results summarized in Table II show that the SVM with a string kernel has a better overall performance and the detection system has a good performance (with an overall accuracy above 95%) in detecting malicious spam emails.

Table II. Performance comparison between classifiers on malicious spam detection

CLASSFIER	ACC (%)	PRE (%)	REC (%)	F1
Decision Tree	93.0	86.6	90.11	0.864
Naïve Bayesian	94.5	89.5	91.0	0.889
String SVM	**95.3**	**90.9**	**95.7**	**0.925**
kNN	89.3	84.3	81.9	0.811

５．主な発表論文等

〔雑誌論文〕（計　５　件）

① Jianliang Wu, Tingting Cui, Tao Ban, Shanqing Guo, and Lizhen Cui, PaddyFrog: systematically detecting confused deputy vulnerability in Android applications, Security and Communication Networks, with review, published online, 30 Jan, 2015

② Siti-Hajar-Aminah Ali, Seiichi Ozawa, Junji Nakazato, Tao Ban, Jumpei Shimamura, An online malicious spam email detection system using resource allocating network with locality sensitive hashing, Journal of Intelligent Learning Systems and Applications, with review, Vol. 7, pp.

42-57, 2015

③ Ryoichi Isawa, Tao Ban, Shanqing Guo, Daisuke Inoue, and Koji Nakao, An accurate packer identification method using support vector machine, IEICE Transactions on Fundamentals, with review, Vol. 97, No. 1, pp. 153-263, 2014

④ Shaoning Pang, Fan Liu, Youki Kadobayashi, Tao Ban, and Daisuke Inoue, A learner-independent knowledge transfer approach to multi-task learning, Cognitive Computation, with review, Vol. 6, No. 3, pp. 304-320, 2014

⑤ Ajit Narayanan, Yi Chen, Shaoning Pang, and Tao Ban, The effects of different representations on static structure analysis of computer malware signatures, The Scientific World Journal, with review, Vol. 2013, Article ID 671096, 2013

〔学会発表〕（計 10 件）

① Aminah Ali Siti Hajar, Seiichi Ozawa, Junji Nakazato, Tao Ban, and Jumpei Shimamura, An autonomous online malicious spam mail detection system using extended RBF network, Accepted by IJCNN 2015, with review, Killarney, Ireland, July 14, 2015

② Tao Ban, Masashi Eto, Shanqing Guo, Daisuke Inoue, Koji Nakao, and Runhe Huang, A study on association rule mining of darknet big data, Accepted by IJCNN 2015, with review, Killarney, Ireland, July 14, 2015

③ Shaoning Pang, Yiming Peng, Tao Ban, Daisuke Inoue, and Abdolhossein Sarrafzadeh, A federated network online network traffics analysis engine for cybersecurity, Accepted by IJCNN 2015, with review, Killarney, Ireland, July 14, 2015

④ Tao Ban, Masashi Eto, Shanqing Guo, Daisuke Inoue, Koji Nakao, and Shaoning Pang, Association rule mining for big darknet traffic data, 13th International Conference on Neuro-Computing and Evolving Intelligence, without review, Auckland, New Zealand, February 20, 2015

⑤ Yuli Dai, Shunsuke Tada, Tao Ban, Junji Nakazato, Jumpei Shimamura, Seiichi Ozawa, Detecting malicious spam mails, an online machine learning approach, ICONIP 2014, with review, pp. 365-372, Kuching, Malaysia, October 31, 2014

⑥ Shaoning Pang, Jianbei An, Jing Zhao, Xiaosong Li, Tao Ban, Daisuke Inoue, and Abdolhossein Sarrafzadeh, Smart task orderings for active online multitask learning, SIAM International Conference on Data Mining, Workshop on Heterogeneous Learning, with review, Philadelphia, Pennsylvania, USA, April 26, 2014

⑦ 多田隼介、中里純二、班涛、島村隼平、小澤誠一、スパムメールに対するオンライン悪性度判定システムの開発、暗号と情報セキュリティシンポジウム（SCIS2014）、査読なし、鹿児島県、鹿児島市、城山観光ホテル、2014 年 1 月 24 日

⑧ Tao Ban, Ryoichi Isawa, Shanqing Guo, Daisuke Inoue, and Koji Nakao, Efficient malware packer identification using support vector machines with spectrum kernel, ASIAJCIS 2013, with review, pp. 69-76, Seoul, Korea, July 25, 2013

⑨ Tao Ban, Ryoichi Isawa, Shanqing Guo, Daisuke Inoue, and Koji Nakao, Application of string kernel based support vector machine for malware packer identification, IJCNN 2013, with review, pp. 1-8, Dallas, Texas, USA, August 5, 2013

⑩ 中里純二、班涛、島村隼平、衛藤将史、井上大介、中尾康二、メール転送経路に着目したスパムメール分析、ICSS 研究会、査読なし、沖縄県、名護市、名桜大学、2013 年 3 月 28 日

６．研究組織

(1)研究代表者
班　涛　（BAN, Tao）
独立行政法人情報通信研究機構・ネットワークセキュリティ研究所・サイバーセキュリティ研究室・主任研究員

研究者番号：80462878