Nonlinear Phenomena in Complex Systems, vol. **17**, no. 3 (2014), pp. 284 - 289

Dimension Reduction in Network Attacks Detection Systems

V. V. Platonov and P. O. Semenov
Saint-Petersburg State Polytechnic University,
29 Politechnicheskaya Str, 195251 Saint-Petersburg, RUSSIA

(Received 21 March, 2014)

Possibility of application of dimension reduction techniques to generate a list of most significant parameters for detecting network attacks is analyzed. The model of intrusion detection system with modular architecture is proposed, which allows packages classification by different support vector machines.

AMS Subject Classification: 68U35
Keywords: intrusion detection system, support vector machine, dimension reduction, principal component analysis

1. Introduction

The research is aimed on applicability of intellectual data analysis. Principal component analysis was chosen for the problem solution. Its application resulted in construction of a weight matrix W_{kxp} for calculation of new parameters as $S = W_{k\ldots p}X$. For the purposes of this investigation a program prototype of the intrusion detection system was developed using support vector method (SVM) as a classifier and principal component method (PCM) for signature space formation that are most suitable for each attack under consideration. IP, TCP, UDP headlines and ICMP packages TCP response parameters were used for network packages classification, C language has been used as a programming language. libSVM was used to work with the support vector method (SVM), with DARPA training files being used for investigation. The resulting intrusion detection system has modular architecture: each module is responsible for detection of a certain attack set. A simplified scheme of performance of each detection module in the constructed intrusion detection system (IDS) is shown in Fig. 1. The formation of individual SVM-modules is followed by IDS modular architecture adjustment. The special feature of this approach is division of all considered network attacks into separate groups with their own detection block and detailed adjustment. Signals from individual modules are filtered in the response block, thus selecting only the checked attack signal. This results in a reduced number of the system false responses.

1.1. Automation of detection module adjustment

An algorithm of automatic adjustment selection for the support vector machine block and the dimension reduction block was developed for the programming complex. Fig. 2 shows a simplified algorithm scheme. In the "Basic Parameter Extraction" block headline parameters of network and transportation levels are extracted and TCP sessions parameters are calculated. In the "Dimension Reduction" block "new" parameters for vector sets with basic parameters are calculated by means of the principle component analysis (PCA). By applying δ threshold value negligible basic parameters are discarded, while a list of most significant "new" parameters is formed by means of threshold ξ value. In the "SVM Training" block a SVM model is generated for the vector set with "new" parameters. In the "SVM checking" block a percentage of correctly classified vectors is determined. In the "Work Analysis" block based on the preceding chain blocks the decisions concerning the alteration of the internal module

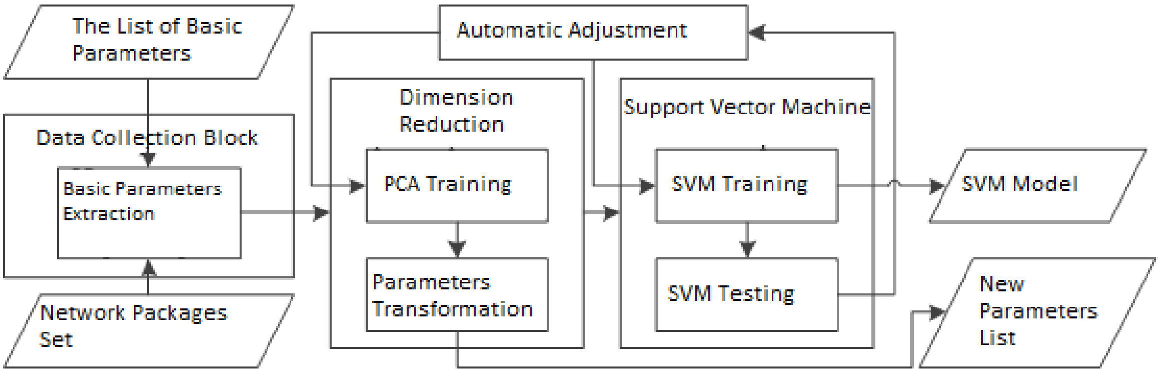

FIG. 1: Intrusion detection module formation.

parameters, algorithms and properties are made. In Table 1 all automatically adjusted parameters for the system main blocks are listed. The purpose of this block is to achieve the highest percentage of correctly classified packages and the system highest performance speed. The parameter selection process consists of three enclosed cycles in Fig. 2. SVM adjustment selection is based on the construction of the "search network" for radial basis core by C and γ parameters. An algorithm of SVM parameter selection for radial basis core by C and γ parameters is suggested by the authors of library LibSVM [1] . When attempting to apply this algorithm for the considered data it was found that the suggested intervals for $\log_2 C$ and $\log_2 \gamma$ are not appropriate. Moreover, with $\log_2 C$ exceeding a certain constant the result of SVM work depends only on γ parameter. In the course of the performed study an automatic algorithm of SVM parameter selection was developed: first, $\log_2 \gamma$ parameters with the least number of support vector in training are determined for $C = 2^{25}$ (selection is performed with an interval $\Delta(\log_2 \gamma) = 2$). The $\log_2 C$ parameter in the area of the best $\log_2 \gamma$ values is found. Next, "search network" with an interval $\Delta(\log_2 C) = 2^{-4}$ and $\Delta(\log_2 \gamma) = 2^{-4}$ is constructed and the best values in the area of best points obtained at the previous algorithm stage are found. The distribution of the best points based on the experimental results for various attacks and basic parameter sets is shown in Fig. 3.

Table 1: Automatically adjusted parameters.

Basic parameters extraction	Dimension reduction	SVM training
Basic parameters list	analyzed matrix	SVM nucleus
parameters order	training set	C parameter
scaling formula	δ threshold parameter	γ nucleus parameter
	ξ threshold parameter	*degree* nucleus parameter
		$coef_0$ nucleus parameter

Hues of grey color indicate the areas with identical detection percentage and support vector number. The darkest area corresponds to the best values set. All points below the first curve have a great number (up to several thousands) of support vectors, with the training process lasting from several minutes to several hours. A dramatic reduction of support vectors number and, accordingly, a dramatic increase of training and testing speed is observed in the area to the right of the second curve. To the right of the third curve the number of support vectors and the correct classification percentage do not change with further increase of C parameter. The best

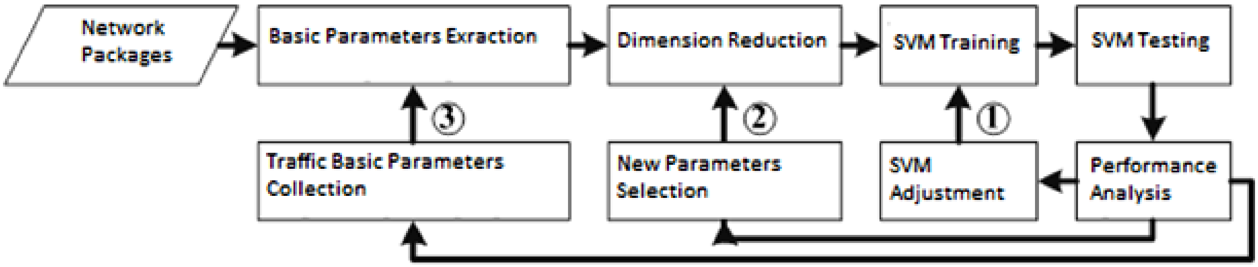

FIG. 2: Automated adjustment process.

points can lie on the horizontal ray starting from curve 3 or at the point of the distinct area between the second and third curves After obtaining the best result for SVM testing (the least number of errors of the first and second types (FN and FP) and the least number of support vectors in SVM model) the automatic adjustment block changes the parameters in dimension reduction block and repeats the same operation cycle in SVM module. To increase the capabilities of the programming complex the feature of principal component training for an attack set only was added. With such an approach the dimension reduction block determines internal dependencies between the parameters within the framework of attack packages rather than those in the entire training set. In the attack set training application of new parameters with the least eigenvalues of the weight matrix is similar to the attack signature It should be noted that such an approach depends considerably on the training set, which can result in an overtraining problem and considerable number of false responses. This approach proves highly efficient for certain attacks (not exceeding 5 new parameters and less than 20 support vectors for 100% recognition), however, it entirely inapplicable for others, which makes this approach only an additional means in the designed programming complex. Automatic adjustment in the dimension reduction block boils down to the selection of the matrix to be analyzed, a set of vectors to be considered (all packages or only attack ones) and the selection of two threshold values of δ and ξ. As a result of the automatic adjustment, block work the system best parameters are found:

- the least FP and FN values determine the attack detection quality;
- the least number of basic parameters, new parameters and support vectors of SVM-model determine the system's performance speed.

1.2. SVM Parameters Selection

Performance of the programming prototype with User-to-Root and Remote-to-local categories from DARPA classification was tested. The following features causing poor recognition were typical for the data under consideration coming into SVM:

- in analyzed data the number of packages with "attack" tag does not exceed 1% of the total packages number, while in the vast majority of current studies devoted to the operation of support vector machines with various data approximately the same amount of vectors is used in each class;
- the second important feature is relatively small dimensions of the data under investigation – 2 dozens parameters at the most vs. 50 or several hundreds in other well-known SVM studies.

Separate attacks with application of network filtration for IP victim address were considered to counteract the first feature. To study the second feature, testing was divided into three parts, depending on basic parameters sets under consideration: multi-order IP and TCP parameters, 8-order IP and TCP headlines parameters and 8-order headline parameters with added TCP session parameters.

2. Multi-order parameters of IP and TCP Packages

Multi-order traffic parameters extracted from IP and TCP packages headlines were used at the first testing stage, with totally 14 basic parameters being used: 6 for IP and 8 for TCP. During the experiments the following results were obtained for various attacks and network dumps:

- For a considerable part of the attacks 100% recognition was achieved with 30-400 vectors;
- for some dumps it was impossible to perform support vector machine training for particular attacks in acceptable time (within1,5-2 hours period), possible cause being inaccuracies in DARPA attack description;
- for the rest of the attacks the recognition result was at least 95%;
- principal component method training and support vectors machine testing was performed for different dumps, with slight changes in results in case of dumps shifting

2.1. 8-order parameters of IP and TCP packages

One of the main features of data testing analyzed in the first part is few parameters for support vector machine operation. Due to this, in order to form the dividing hyper-plane the program has to operate with lower order of the calculated parameters, which results in incapability of SVM training. In order to solve this problem a division of multi-order parameters into several 8-bit parts was introduced into the programming complex (e.g. TCP sequence number can be represented by one, two or four parameters). As a result the maximum number of basic parameters amounted to 24. The experiments similar to the multi-order parameters were performed for the new list of parameters. The following pattern emerged:

- The best results are achieved with least γ, parameter values, with $\log_2 \gamma$ being within the range suggested by libSVM authors;
- for some attacks C parameter was reduced, however, for most attacks it remained considerably greater than that suggested by libSVM library authors;
- for most attacks and dumps for which multi-order parameter training was not carried out or 100% recognition was not achieved , 10% result was obtained for 8-order parameters with the support vectors number not exceeding 300;
- the number of support vectors increased compared to multi-order parameters, which resulted from space dimension increase;
- unlike multi-order parameters the matrix selection plays a considerable part in the principal component method. Operation with the following matrices is possible in the programming prototype: correlation matrix, co-variation matrix, square sum matrix and mixed product matrix;
- for some attacks the least number of support vectors is obtained in SVM-models during operation of the support vector machine with unreduced dimension data.

2.2. TCP-sessions parameters

At the third testing stage, statistical parameters obtained from TCP sessions were included, namely: connection time, the number of sent and received pack-ages, the number of bytes

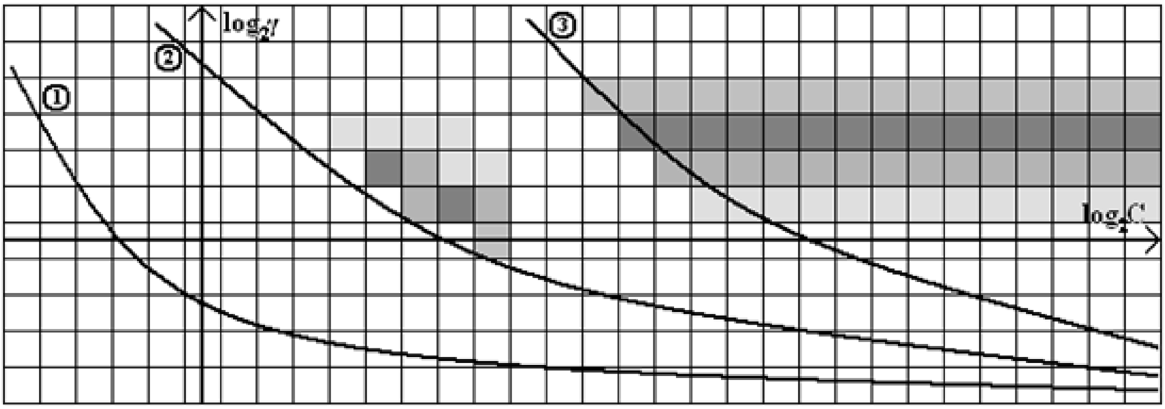

FIG. 3: Search for C and γ parameters for the best SVM classification (RBF core).

and packages with different TCP flags, totaling 49 basic parameters considered. The number of support vectors in SVM-models for network packages classification considerably increased for all analyzed attacks. However, 100% result was impossible to achieve for some attacks at any of the three testing stages. Currently it was found that for some attacks 2 to 5 new parameters out of 49 are sufficient in order to achieve 100% recognition and slight increase of support vectors number. For many attacks a monotonous growth of support vectors number is observed with new parameters number reduction; moreover, for such attacks the least support vector number corresponds to unreduced dimensions operation.

3. Conclusion

The experiments performed with the program prototype show robustness of intrusion detection system and applicability of the selected intelligent data analysis methods for the specified purpose. The support vector machine allows to identify a considerable part of attacks under investigation with 100% confidence, with an error for the rest not exceeding several per cent of the total packages number. The dimension reduction method allows reducing the information volume required for network packages classification and considerably increases the system efficiency. In modular architecture of the intrusion detection system while designing several classifiers of the same packages type, the dimension reduction methods allow to construct pretty simple SVM models that identify intrusions much faster and more accurately than a single general classifier with a great number of parameters and support vectors in a SVM model. The application of various methods, possibility to adjust internal parameters, threshold values enables one to obtain the best possible correlation between the system efficiency and intrusion detection accuracy.

References

[1] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin. A Practical Guide to Support Vector Classification. (National Taiwan University, 2010).

[2] G. Meera Gandhi, Kumaravel Appavoo, S.K. Srivatsa. Effective Network Intrusion Detection using Classifiers Decision Trees and Decision rulesв. Int. J. Advanced Networking and

Applications. **2**, no. 3, (2010).

[3] R. Aarthy and P. Marikkannu. Extended security for intrusion detection system using data cleaning in large database Int. J. Communications and Engineering. **2**, no. 2, 56-60 (2012).

[4] Guy Helmer, J. S.K. Wong, Vasant Honvar, Les Miller, Yanxin Wang. Lightweight agents for intrusion detection. J. Systems and Software. **67**, no. 2, 109-122 (2010).

[5] A. Sperotto, G. Schaffrath, R. Sadre, C. Morariu, A. Pras, B. Stiller. *An Overview of IP Flow-Based Intrusion Detection IEEE communications surveys & tutorials.* Vol. 12, no. 3 (2010).

[6] M. Tavallaee, E. Bagheri, W. Lu, A. A. Ghorbani. A Detailed Analysis of the KDD CUP 99 Data Setв. In: *IEEE Symposium on computational intelligence in security and defence application, 2009.*