Towards the Reduction of Data Used for the Classification of Network Flows

Maciej Grzenda

Warsaw University of Technology,
Faculty of Mathematics and Information Science,
00-661 Warszawa, Pl. Politechniki 1, Poland
M.Grzenda@mini.pw.edu.pl
Orange Labs Poland
02-691 Warszawa, ul. Obrzeżna 7, Poland
Maciej.Grzenda@telekomunikacja.pl

Abstract. The ever growing volume of network traffic results in the need for even more efficient data processing in Intrusion Detection Systems. In particular, the raw network data has to be transformed and largely reduced to be processed by data mining models.

The primary objective of this work is to control the dimensionality reduction (DR) of network flow records in view of the accuracy of misuse detection. A real data set, containing flow records with potential spam messages, is used to perform the tests of the proposed method. The algorithm proposed in this study is applied to investigate the merits of hybrid models composed of dimensionality reduction, neural networks, and decision trees. The benefits of dimensionality reduction and the impact of the process on the overall spam detection rates and false positive rates are investigated. The advantages of the proposed technique over standard a priori selection of reduced dimension are discussed.

Keywords: Network flows, spam detection, dimensionality reduction, multilayer perceptron, decision tree.

1 Introduction

The rapid growth of the Internet results in unprecedented number and scale of security threats. Intrusion detection systems (IDS) [1,3] have been developed in turn to protect the availability, confidentiality and integrity of the network systems [3] and prevent unauthorised use of system resources. Both theoretical models of intrusion detection and reaction [4] and real systems such as SNORT, classified as intrusion prevention/detection system (IPS/IDS)[18] are developed. As the nature of threats dynamically evolves, statistical and artificial intelligence techniques are applied to develop adaptive intrusion detection systems. Such systems are developed at two levels, namely host (Host-Based IDS)[5] and network level (Network IDS, NIDS) [16,20].

E. Corchado et al. (Eds.): HAIS 2012, Part II, LNCS 7209, pp. 68–77, 2012.

Large amounts of data and significant number of threats processed by NIDS resulted in the extensive use of artificial intelligence (AI) techniques in this field. Among them hybrid artificial intelligence systems (HAIS) have been shown to be of particular use. In particular, in [10] a combination of genetic algorithms and fuzzy systems was used to optimise the layout of distributed network sensors used to monitor network traffic. A related idea of a hybrid system combining statistical methods and Petri networks to process data from detection sensors is proposed in [2]. Another approach is based on the idea of combining conventional intrusion detection methods with AI techniques. Among other, the solution proposed in [6] combines immune-inspired methods with conventional network intrusion detection and monitoring methods.

As far as Network Intrusion Detection Systems are concerned, one of the main issues to address is the constantly growing volume of data transferred in modern networks. In the case of TCP connections, each connection is precisely defined by a sequence of datagrams. Formally, the sequence of datagrams of a connection could be represented by a vector x. However, it would be virtually impossible to develop some of the models, such as multilayer perceptrons (MLPs) [9] with thousands or even millions of inputs. Hence, numerous methods of network traffic representation were proposed. Among these methods, the aggregate features of logical network flows, e.g. of TCP connections were proposed. These include the set of 249 discriminators proposed by A. Moore et al. in [17]. The discriminators include a number of features calculated based on the actual sequence of datagrams contained in the x vector, such as maximum of bytes in (Ethernet) packet [17]. From the data processing point of view, the variable dimensionality of the raw traffic data is replaced with a globally constant dimensionality of the flow vectors describing every flow in the network, including TCP and UDP, but also other protocols. It has been shown that some data processing tasks can be successfully performed based on the flow records defined in this way. Among others, H. Kim et al. [12] experimentally showed that flow features-based classification can be successfully used to categorise traffic by service (e.g. http, ftp, smtp).

While flow feature records can be used as an input for data mining models, the impact of individual features on misuse detection remains largely unclear. Hence, in IDS, both feature selection [12,11,20] and dimensionality reduction (DR) [15,22,21] are applied to limit the number of input signals fed to the models. In the case of flow-based detection, the number of flow features can exceed 200, in case most features proposed by A. Moore are applied. Hence, the need for reducing the number of model inputs is also observed.

At the same time, the scale of dimensionality reduction of features proposed in [17] or used in [12] is not discussed in these works and remains largely an open issue. In the case of prediction models, it has been observed that the scale of dimensionality reduction should be controlled by the quality of prediction models developed with the data. In particular, it was shown that a priori data reduction may result in suboptimal prediction models [8]. This study proposes a hybrid AI method combining conventional methods with artificial intelligence

techniques. More precisely, the process of dimensionality reduction is combined with neural networks and decision trees and controlled by the accuracy of misuse detection. A method controlling the scale of dimensionality reduction in view of the accuracy of classification models built with the data transformed by DR is proposed. Simulation results show that the accuracy of the proposed method exceeds the accuracy attained when standard DR techniques, such as widely used [13,14,15] scree plot or proportion of explained variation, are applied to control DR process.

One more outstanding issue is the reduction of false positive (FP) errors [4,1] and the evaluation of the impact of DR on FP rates. Therefore, the objective of this study is to reconsider the existing techniques of selecting reduced dimension of the data. This should be done in view of the compromise between minimal overall and FP classification rates, and maintaining possibly minimal set of input features supplied to the classification model. This formulates a difficult multiobjective optimisation problem this study is aiming to contribute to.

One of the factors hindering the research on Internet vulnerabilities is lack of realistic and representative data sets [7]. To address the objectives formulated above, our works relies on the fundamental work of M. Žádník and Z. Michlovský [19]. The authors developed the real network flow data set to analyse one of the key vulnerabilities of Internet systems i.e. spam messages. The data set made available by the authors, is discussed in detailed in Sect. 2. The flow records labelled with spam category enable research into the accuracy of net flow-based misuse detection.

In order to analyse the impact of DR on network flow classification, hybrid classifiers based on the combination of DR, decision trees and MLPs were developed. The impact of reduced dimensions on the spam classifiers is evaluated through the extensive battery of simulations.

The remainder of this paper is organised as follows:

- Sect. 2 presents the problem and the data set used in the experiments,
- Sect. 3 discusses the proposed methods and the way they are used to deal with network flows and spam detection,
- Next, results are discussed in Sect. 4, which is followed by the conclusions outlined in Sect. 5.

2 Flow Features-Based Spam Detection

2.1 The Spam Data Set

This study follows the fundamental work of M. Žádník and Z. Michlovský [19]. The authors captured network traces for the communication with the mail server, created the net flow records from the traces and labelled every flow record with an appropriate class. The delivered mails were classified by SpamAssassin into two groups: relevant emails and spam [19]. The details of the procedure can be found in [19]. The data set contains 58 042 records, with 64 non-constant input features each. The number of records in every class is different, which is

Table 1. Network flow categories

Class name	No. of records	Class description
y_spam	11222	Spam messages, as classified by SpamAssassin
n_spam	1554	Valid messages
dnsbl	38314	Spam messages, rejected based on server black list
relay	2618	Outgoing messages, sent from the monitored mail server
other	4334	Traffic caused by scanning, DoS, etc.

summarised in Table 1 [19]. The results of the work and the data set developed by the authors were used as a starting point for this study.

2.2 Data Preprocessing

To accommodate the training process, a more balanced data set can be created. Due to relatively high number of records in every class, a decision was made to first apply undersampling to the original data set. Therefore, the number of records in every class, was reduced by random selection to $card(\{x : class(x) = n\_spam\})$ i.e. to 1554. Moreover, from the network point of view, it is important to classify network flows into *spam* flows and *correct* flows, irrespective of the subcategories of this division. Finally, the third class meaning *other* can be used to contain other traffic such as traffic caused by scanning. Hence, a decision was made to reduce the number of classes. In particular, one *spam* class contains both *dnsbl* and *y_spam* original classes. Similarly the *correct* class contains both *relay* and *n_spam* classes. For the purpose of the remaining part of this study, the three classes: *spam*, *correct* and *other* are used to mark all the patterns.

3 Model Development and Data Reduction

3.1 Key Objectives

The data set described above provides basis for the construction of classifier models. Moreover, under laboratory conditions, flow records of arbitrary sizes can be considered and processed efficiently. However, in modern telecom networks, the observed network traffic can exceed 10Gbit/s. This makes the need for high performance data collection and investigation even more significant than before. Hence, the error rates of a classification model, should be analysed also in terms of the computing overhead of both the model and the data capture module producing flow records out of raw network data. It should be emphasised that both operations i.e. flow record production and flow classification should be performed in near real-time conditions. Hence, in the analysed case of network flow-based classification, an important aspect is to minimise the length of a flow record required to classify every flow. The impact of this reduction on the classification accuracy might be different, depending on the category of classification

model. Hence, one of the objectives of model development and simulation scenarios is to experimentally evaluate the impact of flow data reduction on different classification models.

3.2 The Evaluation Method

The analysis of high-dimensional data aims to identify and eliminate redundancies among the observed variables [15]. The process of DR is expected to reveal underlying latent variables. More formally, the DR of a data set $D \subset \mathbb{R}^S$ can be defined by a function used to code an element $x \in D$ [15]:

$$c : \mathbb{R}^S \longrightarrow \mathbb{R}^R,\ x \longrightarrow \tilde{x} = c(x) \tag{1}$$

In the analysed case of the dimensionality reduction of flow records D the primary objective is to minimise the reduced record length R, while maintaining possibly high classification accuracy and relatively limited false positive error rates. To objectively investigate the impact of the reduction on the spam detection classifier, Alg. 1 was used. The algorithm is prepared to work with the data sets composed of a relatively limited number of records. More precisely, Cross-Validation (CV) is used in $EvaluateDimension()$ procedure to evaluate the merits of the data transformation.

Input: D - matrix of input attributes, $P \subset \mathbb{R}$ - vector of corresponding output features, $card(D)$=$card(P)$, K - the number of CV folders, r - the number of training sessions
begin
  for $R = 2, \ldots, S$ **do**
    $[E^R_{\mathrm{V}}(D,P), F^R_{\mathrm{V}}(D,P), E^R_{\mathrm{T}}(D,P), F^R_{\mathrm{T}}(D,P)] =$ EvaluateDimension(D,P,K,R,r);
  end
end

Algorithm 1. The evaluation of the impact of dimensionality reduction on classifier accuracy

The core of the proposed method is the evaluation of the impact of dimensionality reduction to dimension R on the classification model built with $c(D) \subset \mathbb{R}^R$. This evaluation is described by Alg. 2 and considers different divisions of the available data set and several runs of model construction algorithm to minimise the impact of individual sessions on dimension evaluation. Moreover, both $E()$ rate standing for classification error and $F()$ standing for false positive rates are calculated. In the analysed case, by false positive a classification of a non-spam flow i.e. *correct* or *other* flow, to *spam* class is considered. The algorithms automate the task of evaluating the impact of dimensionality reduction method defined by $DimensionalityMapping()$ on the classification models.

Input: D - matrix of input attributes, $P \subset \mathbb{R}$ - vector of corresponding output features, $card(D)$=$card(P)$, K - the number of CV folders, R - reduced dimension, r - the number of training sessions
Data: D_i - a family of K sets: $D_i \cap D_j = \emptyset, i \neq j$, $\cup_{i=1}^{K} D_i = D$; $P_i, P_{\mathrm{L}}, P_{\mathrm{T}}, P_{\mathrm{V}}$ - output features corresp. to $D_i, D_{\mathrm{L}}, D_{\mathrm{T}}, D_{\mathrm{V}}$
Result: $E_{\mathrm{T}}^{R}(D,P)$ - the average classification error rate of the models on the testing sets; $E_{\mathrm{V}}^{R}(D,P)$ - the average classification error rate of the models on the validation sets; $F_{\mathrm{T}}^{R}(D,P)$ - the average false positive error rate of the models on the testing sets; $F_{\mathrm{V}}^{R}(D,P)$ - the average false positive error rate of the models on the validation sets.

begin
  c = DimensionalityMapping(D,R);
  for $i = 1, \ldots, r$ **do**
    $\{D_j, P_j\}_{j=1,\ldots,K}$ = DivideSet(D,P,K);
    for $k = 1 \ldots K$ **do**
      $D_{\mathrm{T}} = D_k$;
      $D_{\mathrm{V}} = D_{(k+1) \bmod K}$;
      $D_{\mathrm{L}} = \cup_{j \in \{1,\ldots,K\} - \{k\} - \{(k+1) \bmod K\}} D_j$;
      $\tilde{D}_{\mathrm{L}}^{R} = c(D_{\mathrm{L}})$;
      $\tilde{D}_{\mathrm{T}}^{R} = c(D_{\mathrm{T}})$;
      $\tilde{D}_{\mathrm{V}}^{R} = c(D_{\mathrm{V}})$;
      $M = train(\tilde{D}_{\mathrm{L}}^{R}, P_{\mathrm{L}}, \tilde{D}_{\mathrm{V}}^{R}, P_{\mathrm{V}})$;
      $E_{\mathrm{T}}((i-1)*K+k) = E(M(\tilde{D}_{\mathrm{T}}^{R}), P_{\mathrm{T}})$;
      $E_{\mathrm{V}}((i-1)*K+k) = E(M(\tilde{D}_{\mathrm{V}}^{R}), P_{\mathrm{V}})$;
      $F_{\mathrm{T}}((i-1)*K+k) = F(M(\tilde{D}_{\mathrm{T}}^{R}), P_{\mathrm{T}})$;
      $F_{\mathrm{V}}((i-1)*K+k) = F(M(\tilde{D}_{\mathrm{V}}^{R}), P_{\mathrm{V}})$;
    end
  end
  $E_{\mathrm{T}}^{R}(D,P) = median(E_{\mathrm{T}}())$;
  $E_{\mathrm{V}}^{R}(D,P) = median(E_{\mathrm{V}}())$;
  $F_{\mathrm{T}}^{R}(D,P) = median(F_{\mathrm{T}}())$;
  $F_{\mathrm{V}}^{R}(D,P) = median(F_{\mathrm{V}}())$;
end

Algorithm 2. EvaluateDimension() procedure

4 Experimental Results

The experiments performed in this study involved decision trees and multilayer perceptrons as classification models. Separate runs of Alg. 1 were performed for these two types of models. In both cases, the *train*() method being a part of Alg. 2 was used to train a model being a decision tree or an MLP network using training $(\tilde{D}_{\mathrm{L}}^{R}, P_{\mathrm{L}})$ and validation $(\tilde{D}_{\mathrm{V}}^{R}, P_{\mathrm{V}})$ data sets. Moreover, it was aiming to

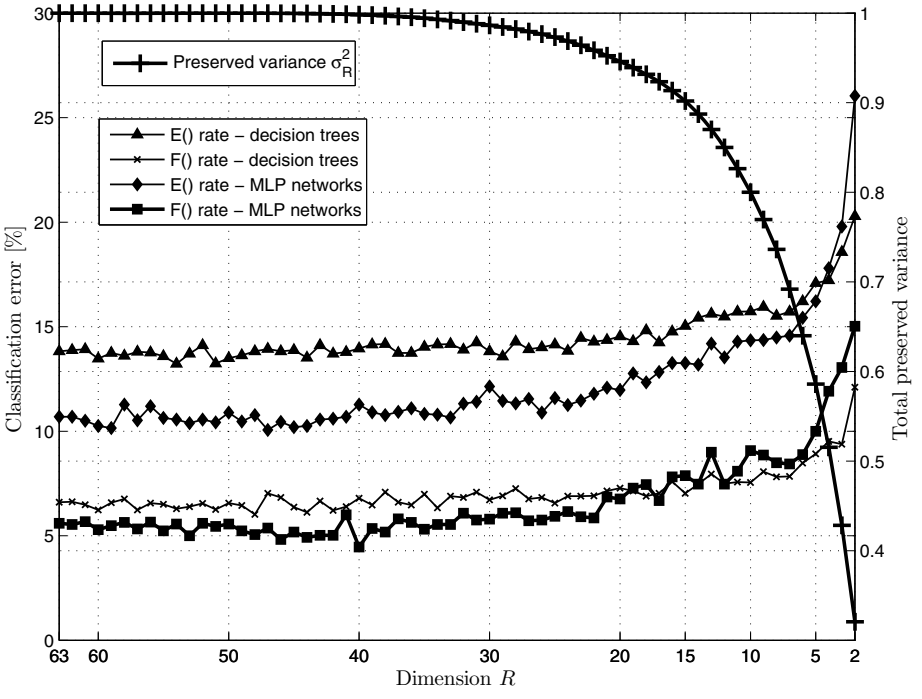

Fig. 1. Spam detection error rates as a function of reduced dimension R

improve the generalisation capabilities of a model. In the case of decision tress, first the construction of an excessive tree with minimum split criterion set to 1 record was done. Then, the tree was reduced by eliminating unnecessary leaves and divisions, in order to minimise the classification error rate measured on the validation data set. In the case of MLP networks, early stopping based on the minimisation of the validation error was applied to avoid overtraining.

As far as dimensionality reduction is concerned, one of the most popular techniques, being Principal Component Analysis (PCA) [15] was used. Hence, the role of $DimensionalityMapping()$ was to return an $R \times card(D)$ PCA transformation matrix that was used next to reduce the dimensionality of the original data.

Let ϱ denote a $(S \times S)$ covariance matrix, $\lambda_1, \ldots, \lambda_S$ denote eigenvalues of the matrix sorted in descending order and $\mathbf{q}_1, \ldots, \mathbf{q}_S$ the associated eigenvectors i.e. $\varrho \mathbf{q}_i = \lambda_i \mathbf{q}_i, i = 1, 2, \ldots, S$. Finally, let $\mathbf{a} : a_j = \mathbf{q}_j^T \mathbf{x} = \mathbf{x}^T \mathbf{q}_j,\ j = 1, \ldots, S$. The coding function $c^R(\mathbf{x})$ reducing dimension to R was defined as $c^R(\mathbf{x}) = [a_1, \ldots, a_R]$.

Extensive simulations were performed to analyse the error rates $E()$ and $F()$ over a range of reduced dimensions $R \in \{2, ..., 63\}$, with $K = 3$ and $r = 4$. In the case of MLP networks, 10 hidden neurons were used. The results of the experiments are summarised in Fig. 1. The error rates on the testing data sets

Table 2. The summary of the search for the optimal dimension R

		E() - overall classification		F() - false positive rate	
R_{SP}	R_{PEV}	R_{MLP}	R_{DT}	R_{MLP}	R_{DT}
35	21	30-33	18-20	22-24	15-18

calculated by Alg. 1 are shown together with the Proportion of Explained Variation (PEV) value corresponding to the analysed dimension R.

As both decision tree creation and the training of MLP networks is partly affected by different divisions of the entire data set and the nature of the model construction process, which may produce different models, the resulting $E()$ and $F()$ curves are partly noisy. Nevertheless, the following conclusions can be made:

- The flow records can be largely reduced from the original $S = 64$ dimensions. A reduction to ca. 20 signals can largely preserve the error rate levels observed at $R = 63$.
- Should higher data reduction be needed due to performance reasons, this may impact the model selection. More precisely, MLP networks outperform decision trees for $R \geq 5$, when overall rate $E()$ is considered, while the opposite tendency is observed for $R < 5$.
- The selection of the best classification model depends on the category of error rate to minimise. In case, the minimisation of false positive errors is the dominating criterion and $5 < R < 15$, decision trees provide lower error rates. For the same dimensionality of the data, MLP networks provide lower overall classification error.

Diverse values of optimal reduced dimension R selected by different techniques are summarised in Table 2. R_{SP} and R_{PEV} stand for the reduced dimension as suggested by scree plot analysis and PEV criterion. In the latter case $R_{\mathrm{PEV}} = min_{l=1,2,\ldots,S} : \frac{\sum_{d=1}^{l} \lambda_d}{\sum_{d=1}^{S} \lambda_d} \geq 0.95$. The remaining values R_{MLP} and R_{DT} show the optimal dimensions when a compromise between the minimisation of $E()$ and $F()$ and the minimisation of R is sought, while using MLP networks and decision trees, respectively. What should be emphasised is that the a priori selection of reduced dimension $R \in \{R_{\mathrm{SP}}, R_{\mathrm{PEV}}\}$ may result in suboptimal R values. Equally importantly, the optimal dimension is largely different depending on the classification model. For MLP networks, when overall classification accuracy is the dominating criterion, $R \in \{30, \ldots, 33\}$ should be preferred, while the use of DT results in $R \in \{18, 19, 20\}$. Moreover, in both cases the suggested values are not the values resulting from scree plot analysis.

5 Summary

A method controlling the selection of reduced dimension in view of large scale network data processing was proposed. It was shown that the optimal reduced

dimension depends not only on the a priori investigation of input matrix, but also on the classifier used and the error rate to be minimised. The experiments show that it is possible to classify network flows with relatively high accuracy by using a very limited number of features. What should be emphasised is that when the reduction of false positives is an issue, the scale of dimensionality reduction might be different comparing to the way the same process should be performed when an overall error rate is an issue. Taking into account the excessive volume of data that has to be constantly analysed for possible misuse of the internet services, extensive experiments aiming at selecting the appropriate reduction of the data are fully justified.

Future works should concentrate on the reduction of false positive errors and the reduction of computational cost of the data transformation by eliminating the input attributes having a minor impact on the reduced features. Moreover, other dimensionality reduction techniques are planned to be included in the proposed framework.

Acknowledgments. We would like to thank R. Filasiak for his suggestions and M. Žádník for sharing his experience and the data set used to develop the models discussed in this study.

References

1. Abouabdalla, O., et al.: False Positive Reduction in Intrusion Detection System: A Survey. In: Proc. of IC-BNM 2009, pp. 463–466 (2009)
2. Baláž, A., Trelová, J., Kostráb, M.: Architecture of Distributed Intrusion Detection System Based on Anomalies. In: 14th International Conference on Intelligent Engineering Systems (INES), pp. 79–83 (2010)
3. Barapatre, P., et al.: Training MLP Neural Network to Reduce False Alerts in IDS. In: Proc. of the 2008 Int. Conf. on Computing, Communication and Networking, ICCCN 2008 (2008)
4. Biskup, J.: Security in Computing Systems. Challenges, Approaches and Solutions. Springer, Heidelberg (2009)
5. Dash, S.K., Rawat, S., Pujari, A.K.: Use of Dimensionality Reduction for Intrusion Detection. In: McDaniel, P., Gupta, S.K. (eds.) ICISS 2007. LNCS, vol. 4812, pp. 306–320. Springer, Heidelberg (2007)
6. Fanelli, R.L.: A Hybrid Model for Immune Inspired Network Intrusion Detection. In: Bentley, P.J., Lee, D., Jung, S. (eds.) ICARIS 2008. LNCS, vol. 5132, pp. 107–118. Springer, Heidelberg (2008)
7. Fomenkov, M., Claffy, K.: Internet measurement data management challenges. In: The Cooperative Association for Internet Data Analysis (CAIDA), San Diego, USA (2011)
8. Grzenda, M.: Prediction-Oriented Dimensionality Reduction of Industrial Data Sets. In: Mehrotra, K.G., Mohan, C.K., Oh, J.C., Varshney, P.K., Ali, M. (eds.) IEA/AIE 2011, Part I. LNCS (LNAI), vol. 6703, pp. 232–241. Springer, Heidelberg (2011)
9. Haykin, S.: Neural Networks: a Comprehensive Foundation. Prentice-Hall Inc. (1999)

10. Hu, C., et al.: On the Deployment Strategy of Distributed Network Security Sensors. In: 13th IEEE International Conference on Networks (2005)
11. El-Khatib, K.: Impact of Feature Reduction of the Efficiency of Wireless Intrusion Detection Systems. IEEE Trans. on Parallel and Distributed Systems 21(8), 1143–1149 (2010)
12. Kim, H., et al.: Internet Traffic Classification Demystified: Myths, Caveats, and the Best Practices. In: Proc. of ACM CoNEXT 2008 (December 2008)
13. Larose, D.T.: Data Mining Methods and Models (2006)
14. Lattin, J.M., Carroll, J.D., Green, P.E.: Analyzing Multivariate Data (2003)
15. Lee, J., Verleysen, M.: Nonlinear Dimensionality Reduction. Springer, Heidelberg (2010)
16. Lim, S.Y., Jones, A.: Network Anomaly Detection System: The State of Art of Network Behaviour Analysis. In: Int. Conf. on Convergence and Hybrid Information Technology, pp. 459–465 (2008)
17. Moore, A., Zuev, D., Crogan, M.: Discriminators for use in flow-based classification. Technical Report, RR-05-13, Department of Computer Science, Queen Mary, University of London (2005)
18. `http://www.snort.org/`
19. Žádník, M., Michlovský, Z.: Is Spam Visible in Flow-Level Statistics? CESNET National Research and Education Network, Prague, Czech Republic, Technical Report 6/2008, 67–78 (2008)
20. Zhang, J., Zulkernine, M., Haque, A.: Random-Forests-Based Network Intrusion Detection Systems. IEEE Trans. on Systems, Man, and Cybernetics 38(5), 649–659 (2008)
21. Zhou, Y.-P.: Hybrid Model Based on Artificial Immune System and PCA Neural Networks for Intrusion Detection. In: Proc. of 2009 Asia-Pacific Conf. on Information Processing, pp. 21–24 (2009)
22. Yanwei, F., Yingying, Z., Haiyang, Y.: Study of Neural Network Technologies in Intrusion Detection Systems. In: Proc. of the 5th Int. Conf. on Wireless Communications, Networking and Mobile Computing (2009)