Network Intrusion Detection Based on LDA for Payload Feature Selection

Zhiyuan Tan, Aruna Jamdagni, Xiangjian He, Priyadarsi Nanda
Centre for Innovation in IT Services and Applications (iNEXT)
University of Technology, Sydney, Australia
{Zhiyuan.Tan, Aruna.Jamdagni}@student.uts.edu.au, {xiangjian.he, Priyadarsi.Nanda}@uts.edu.au

Abstract—**Anomaly Intrusion Detection System (IDS) is a statistical based network IDS which can detect attack variants and novel attacks without a priori knowledge. Current anomaly IDSs are inefficient for real-time detection because of their complex computation. This paper proposes a novel approach to reduce the heavy computational cost of an anomaly IDS. Linear Discriminant Analysis (LDA) and difference distance map are used for selection of significant features. This approach is able to transform high-dimensional feature vectors into a low-dimensional domain. The similarity between new incoming packets and a normal profile is determined using Euclidean distance on the simple, low-dimensional feature domain. The final decision will be made according to a pre-calculated threshold to differentiate normal and abnormal network packets. The proposed approach is evaluated using DARPA 1999 IDS dataset.**

Keywords-linear discriminant analysis; feature selection; packet payload; network intrusion detection; Euclidean distance

I. INTRODUCTION

As the popularity of computer network, increasing numbers of transactions have been relocated to network environment. Although this has provided convenience and flexibility to business and human daily life, it has also provided a platform for network criminals and attacks. It was reported that 32,956 vulnerabilities were found through 1995 until the first quarter of 2007 [1], and there was at least one new attack spotted every hours [2]. Therefore, network security has been widely concerned by both industry and research sectors.

To prevent a system from being compromised, Intrusion Detection System (IDS) has become an active research area since it was first introduced in 1980s [3]. Despite advances in intrusion detection, the existing computing infrastructures are still vulnerable to network attacks. This is partially because the most widely used commercial IDSs are misuse-based systems [4][5] and are configured with known attack signatures. They are only effective to the malicious behaviors with the known attack patterns and easy to be evaded by novel attacks. Moreover, it is difficult to keep intrusion detection signature sets updated due to the increasing number of continuously discovered vulnerabilities.

Comparatively, anomaly-based IDSs [6][7] has been proven to be more promising for detection of novel attacks and attack variations. The systems learn normal network traffic behaviors from a set of training data and develop a profile for the normal behaviors. Any deviation of the incoming event profile with respect to the normal profile is considered as anomaly. However, these systems have relatively higher rates of false positives. Thus, in recent years, many studies as shown in [8][9][10][11][12] have been conducted to reduce the high false positive rates, but these approaches have introduced complex, statistical computation in both training and detection phases and have hence caused heavy consumption on system resources and computational power. Therefore, feature reduction becomes essential to create an effective anomaly-based IDS when taking into account the computational complexity and classification performance.

There have been various feature reduction techniques, such as Correlation-based Feature Selection (CFS), Support Vector Machine (SVM), Principal Component Analysis (PCA), Independent Component Analysis (ICA), PCA-ICA, Generalized Discriminant Analysis (GDA) and Linear Discriminant Analysis (LDA), that were discussed and proposed [13][14][15][16][17][18] to reduce the header features of packets. However, there are very few papers that have considered feature selection according to application-layer payload.

The early feature reduction approach [19] on payload, developed by Krugel et al., grouped the byte frequency distributions of 256 ASCII characters into six bins, namely 0, 1-3, 4-6, 7-11, 12-15 and 16-255. Wang et al. [20] proposed an Anagram detector, in which Bloom Filter (BF) was used to reduce memory overhead. Nwanze and Summerville proposed a lightweight payload inspection approach [21], where bit-pattern hash functions were employed to map the bytes at the packet payload onto a set of counters which were the selected features used for intrusion detection.

All existing approaches for feature reduction fail to consider one of the important payload characteristics, i.e., the correlations among the payload features (ASCII characters). The characters in malicious packet payloads present different correlations from those in normal packet payloads.

Geometrical Structure Anomaly Detection (GSAD) model was proposed in [22], which considered the character correlation. GSAD is based on a pattern recognition technique used in image processing, and the model analyses the correlations among ASCII characters in packet payload using the Mahalanobis Distance Map (MDM). GSAD model is proven to have good performance for intrusion detection with low false positive rates and high detection rates. However, because GSAD uses 256^2 features to evaluate and looks for intrusive patterns in a network packet, it creates massive computational complexity.

In this paper, we propose an approach using LDA for feature selection. This approach reduces the computational complexity dramatically while retaining the high detection rates. To our best knowledge, LDA has not been considered in other related researches for payload-based feature selection. Furthermore, this approach uses a difference distance map to order the potential features for feature selection.

This paper is structured as follows. Section II briefly describes the basic concepts of LDA. Section III gives a detailed explanation of LDA-based feature selection approach for intrusion detection. In Section IV, experimental results are given and analysed. Section V draws conclusions and future work.

II. Linear Discriminant Analysis

LDA [23] is one of the commonly used dimensionality reduction and data classification techniques and has been applied in human detection [24], face recognition, speech recognition, marketing research, bankruptcy prediction and network intrusion detection [15][17] etc.

Different from PCA, which extracts features that are the most efficient for representation but may not be useful for discrimination, LDA selects an optimal projection matrix to transform a higher dimensional feature domain to a lower dimensional space while preserving the significant information for data classification. We suppose that there is a set of *n* *d*-dimensional samples $\{x_1, \ldots, x_n\}$ assigned to *k* different classes. Each class C_i, where $i = 1, \ldots, k$, has n_i samples. Projection matrix A_r is found to maximize the between-class scatter matrix

$$S_B = \sum_{i=1}^{k} n_i(\mu_i - \mu)(\mu_i - \mu)^T \tag{1}$$

and minimize the within-class scatter matrix

$$S_w = \sum_{i=1}^{k} \sum_{x \in C_i} (x - \mu_i)(x - \mu_i)^T, \tag{2}$$

where μ is the sample mean of the whole sample set denoted by

$$\mu = \frac{1}{n} \sum_{j=1}^{n} x_j \tag{3}$$

and μ_i is the sample mean for class C_i given by

$$\mu_i = \frac{1}{n_i} \sum_{x \in \mathrm{C_i}} x. \tag{4}$$

Thus, the ratio, *J*, between the between-class scatter matrix S_B and the within-class scatter matrix S_W can be easily maximized by the projection matrix A_r.

$$J = \frac{{A_r}^T S_B A_r}{{A_r}^T S_W A_r} \tag{5}$$

Once the above optimization problem is solved, the classification decision can be easily made on the low dimensional feature space by projecting the original feature space onto the optimal projection matrix A_r.

III. LDA Based Intrusion Detection System

In this paper, LDA is used to select significant features from a Mahalanobis Distance Map (MDM), which is generated by the Geometrical Structure Model (GSM) [22], a key component of the GSAD, for each single network packet to explore the correlations among features (ASCII characters) in a packet payload. Then, the final detection process can be fast conducted on a new low-dimensional domain.

To extract the low-dimensional significant features, difference distance maps need to be generated to measure the difference between normal traffic and particular types of attack traffic, such as the difference between each pair of <*Normal, Phf attack*>, <*Normal, Back attack*> and <*Normal, Apache2 attack*>. Afterwards, LDA is employed to select the most signification features for each normal and attack pair based on the pre-generated difference distance maps. Finally, all of the selected features are integrated into a new significant feature set used for normal profile development and malicious behavior detection. The detailed explanation is given in the following subsections.

A. Feature Selection Using LDA

For the selection of the most significant features, labeled training samples are required and randomly chosen from a normal sample set and various attack sample sets. The techniques shown in [22] are applied to generate MDMs using the training samples. D_1^{normal}, …, D_m^{normal} and D_1^{attack}, …, D_n^{attack} denote the MDMs of all (*m*) normal samples and (*n*) attack samples respectively.

1) Difference distance map

In order to discover the difference between the normal and attack samples, a difference distance map is utilized. We calculate the difference at each element (*i, j*), where *i, j* = 0, …, 255, between the MDMs of the normal samples and the attack samples using Equations (6) to (10) below.

$$\bar{d}_{(i,j)}^{normal} = \frac{1}{m} \sum_{k=1}^{m} d_{(i,j)}^{normal,k} \tag{6}$$

$$\bar{d}_{(i,j)}^{attack} = \frac{1}{n} \sum_{k=1}^{n} d_{(i,j)}^{attack,k} \tag{7}$$

$$\sigma_{normal(i,j)}^2 = \frac{1}{m} \sum_{k=1}^{m} \left(d_{(i,j)}^{normal,k} - \bar{d}_{(i,j)}^{normal}\right)^2 \tag{8}$$

$$\sigma_{attack(i,j)}^2 = \frac{1}{n} \sum_{k=1}^{n} \left(d_{(i,j)}^{attack,k} - \bar{d}_{(i,j)}^{attack}\right)^2 \tag{9}$$

$$diff_{(i,j)} = \frac{(\bar{d}_{(i,j)}^{normal} - \bar{d}_{(i,j)}^{attack})^2}{\sigma_{normal(i,j)}^2 + \sigma_{attack(i,j)}^2} \tag{10}$$

In Equations (6) to (10), $d_{(i,j)}^{normal,k}$ stands for the (*i*, *j*)-th element of MDM of the *k*-th normal sample, $d_{(i,j)}^{attack,k}$ stands for the (*i*, *j*)-th element of MDM of the *k*-th attack sample, $\bar{d}_{(i,j)}^{normal}$ and $\sigma_{normal(i,j)}^2$ denote the mean and the variance of the (*i*, *j*)-th elements of the normal sample

MDMs, and $\bar{d}_{(i,j)}^{attack}$ and $\sigma_{attack(i,j)}^{2}$ denote the mean and the variance of the (i, j)-th elements of the attack sample MDMs. The difference at element (i, j) between the normal samples and the attack samples is denoted by $diff_{(i,j)}$and computed by Equation (10). The difference distance map between the normal samples and the attack samples is defined by $Diff = \left[\, diff_{(i,j)}\right]_{256\times256}$. A difference distance map is generated for each pair of normal traffic and particular type of attack traffic, and will be used for the selection of significant features.

Because the dimension of the difference distance map is large, it is very time consuming if the map is directly used to differentiate the normal traffic and the attack traffic. Therefore, we propose to use LDA for feature selection (i.e., to reduce the dimension of the map).

2) LDA-based feature selection

In the difference distance map, the larger a feature (i.e., a matrix element) is, the more important the feature is to discriminate attack traffic from normal traffic. We first select the most significant r features from the difference distance map. The element locations of these features in the difference distance map determine the element locations in every MDM of a normal or an attack sample to form a corresponding r dimensional distance vector represented by $D_{\mathrm{r,k}} = [d_{k(U_{r,1}, V_{r,1})}, d_{k(U_{r,2}, V_{r,2})}, \ldots, d_{k(U_{r,r}, V_{r,r})}]^{\mathrm{T}}$, where $(U_{\mathrm{r},1}, V_{\mathrm{r},1})$, $(U_{\mathrm{r},2}, V_{\mathrm{r},2})$, …, $(U_{\mathrm{r,r}}, V_{\mathrm{r,r}})$ indicate the element locations of the largest r features in the difference distance map, r is ranged from 1 to 256^2 and k indicates the k-th sample. Let $D_{r,k}^{normal}$ and $D_{r,k}^{attack}$represent the $D_{\mathrm{r,k}}$ of the k-th normal sample and the k-th attack sample respectively. Then, the projection vector A_r is computed by

$$A_r = (\sum \overline{D}_r^{normal} + \sum \overline{D}_r^{attack})^{-1}(\overline{D}_r^{normal} - \overline{D}_r^{attack}) \quad (11)$$

where $\overline{D}_r^{normal}$ and $\overline{D}_r^{attack}$ are the averages of $D_{r,k}^{normal}$ and $D_{r,k}^{attack}$, and $\sum \overline{D}_r^{normal}$ and $\sum \overline{D}_r^{attack}$ are the covariances of $D_{r,k}^{normal}$ and $D_{r,k}^{attack}$. The whole process will be conducted iteratively until the number of significant features reaches the pre-set value, and the final projection matrix A_r will be determined. Once the projection vector is finalized, the corresponding final set of features is considered as the most significant features.

The above is the feature selection process in this paper for detection of each type of attack. For all types of attacks, we need to combine the selected features into a new significant feature set, which is used for normal profile development and malicious behavior detection.

B. Normal Profile Development

The normal profile is utilized to detect the similarity between the normal behavior and new incoming packet. It is developed by using the normal training samples and the selected significant feature set. In this section, we explain how to perform the development of the normal profile.

Mean values of the significant r features of all normal training samples and a detection threshold are the basic components of the normal profile. Given a set of normal training samples $X = \{x_1, \ldots, x_m\}$, which have been applied in the feature selection phase, and the significant feature set $F_{\mathrm{k}} = [f_{k(U_1, V_1)}, f_{k(U_2, V_2)}, \ldots, f_{k(U_r, V_r)}]^{\mathrm{T}}$, in which (U_1, V_1), (U_2, V_2), …, $(U_{\mathrm{r}}, V_{\mathrm{r}})$ indicate the locations of the significant r features and k indicates the k-th sample. The mean values are denoted by

$$\overline{F} = \frac{1}{m}\sum_{k=1}^{m} F_k, \quad (12)$$

and they are stored in the normal profile used for comparing with any new incoming packet. Threshold is another important component to consider. Without an appropriate criterion, it is hard to achieve a satisfactory detection performance. The larger the threshold value is, the less false positive alarm is generated. On the other hand, smaller threshold will in turn create a higher detection rate.

In this paper, we select a threshold through a distribution analysis of the Euclidean distance between each normal training sample and the mean value $\overline{F}$. The Euclidean distance from the k-th normal training sample to the mean value $\overline{F}$ is computed by

$$ED_k = \sqrt{\sum_{i=1}^{r}(f_{k(U_i,V_i)} - \overline{f_{(U_\imath,V_\imath)}})^2}. \quad (13)$$

$\overline{f_{(U_\imath,V_\imath)}}$ is the $(U_{\mathrm{i}}, V_{\mathrm{i}})$-th element of $\overline{F}$. The standard deviation of the Euclidean distances from the k-th normal training sample to the mean value $\overline{F_r}$ of the normal training samples is

$$\delta = \sqrt{\frac{1}{m-1}\sum_{k=1}^{m}(ED_k - \overline{ED})^2}\,, \quad (14)$$

where $\overline{ED} = \frac{1}{m}\sum_{k=1}^{m} ED_k$ .We assume that the distance ED_k is of normal distribution, so three standard deviations account for 99% of the sample population.

C. Attack Recognition

Similar to the normal profile development process, for any new incoming packet, the GSM is applied to generate the MDM of the packet. Then, the most significant r features are collected to form a feature vector F from the MDM. Afterwards, the Euclidean distance between F and $\overline{F}$ is calculated using Equation (13). The incoming packet is considered as an attack or a threat if and only if the Euclidean distance from F to $\overline{F}$ is greater than +3δ or smaller than -3δ, where δ is the standard deviation computed by Equation (14).

Compared with the pure GSAD model in [22], the approach in this paper uses only the selected significant features for discriminating the normal traffic from the attack traffic. It avoids the heavy computational complexity without using all 256^2 features in an MDM.

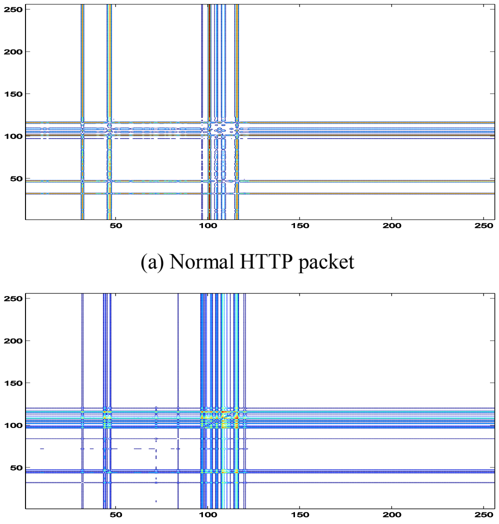
IV. Experimental Results and Analysis

Experimental results are presented and analysed in this section. We evaluate the LDA feature selection approach on the DARPA 1999 IDS dataset [25], which contains five-week network traffic recorded in tcpdump format. Week 1 and week 3 data are attack-free data, and the other three-week data contain both normal and attack network traffic. The DARPA 1999 dataset consists of five main categories of attacks including scan or probe, DoS, R2L, U2R and data.

A. Experimental Results

In our experiments, we consider inbound HTTP traffic only. HTTP-based attacks are mainly from the HTTP GET request at the server side. We use the same conditions as the [22] did. The LDA-based IDS is trained and tested with the inbound HTTP GET request traffic carrying payload extracted on week 4 (5 days) and week 5 (5 days). The extracted HTTP traffic packets corresponding to HTTP service are destined to two different HTTP servers existing in DARPA 1999 dataset: marx (Linux Server with IP address 172.16.114.50) and hume (NT Server with IP address 172.16.112.100). The total numbers of packets after filtering are 783,443 for marx, and 8431 for hume hosts respectively. Then, we further filter the normal and attack HTTP GET request packets, and divide them into normal and attack datasets respectively. We randomly choose 300 normal packets and 900 attack packets for feature selection, and choose another 300 packets for normal model training. Finally, we randomly select 1000 normal packets and 3000 attack packets from the remaining for test. The attack packets contain Apache2 attack, Back attack and Phf attack.

In the feature selection stage, the signification features are selected using the randomly chosen 300 normal packets and 900 attack packets for each type of attack according to the process discussed in Section III A. The normal model of LDA-based IDS is trained on the 300 training packets given in Section III B.



(a) Normal HTTP packet

(b) Phf attack packet

Figure 1. Average Mahalanobis distance maps of normal HTTP and Phf attack packets

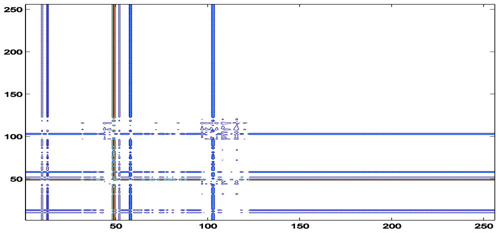


Figure 2. Difference distance map between normal HTTP and Phf attack packets

TABLE I. Confusion Matrix for LDA-based IDS

Predicted Actual	Normal	Attack	%Correct
Normal	963	37	96.3
Attack	0	3000	100
% Correct	100	98.78	

Figs. 1(a) and (b) show the average MDMs of the normal HTTP samples and the Phf attack samples, and the difference distance map is shown in Fig 2. There are totally 256^2 features in each of the average MDMs and the difference distance map. As can be seen from the following figures, those normal and attack samples present clearly different behaviors.

We conduct several experiments to extract the optimal number of significant features to best separate normal packets from attack packets. The optimal result is found to be 100 features selected by LDA for each of three types of attacks. Then, the normal profile is developed based on the combined 300 significant features.

In the test stage, the trained LDA-based IDS is evaluated on the testing dataset containing both the normal packets and the attack packets. The results are shown in Table I. A detailed analysis is given in the next subsection.

B. Experimental Analysis

The results in Table I reveal that the 300 optimally selected significant features can well differentiate the attack packets from the normal packets. In this section, the information contained in the confusion matrix is further analysed using Detection Rate (DR) and False Positive Rate (FPR). We introduce evaluation metrics for the analysis. The metrics used are True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

- *TP*: the number of actual attack classified as attack
- *TN*: the number of actual normal classified as normal
- *FP*: the number of actual normal classified as attack
- *FN:* the number of actual attack classified as normal

The Detection Rate is given by $DR = \frac{TP}{TP+FN} \times 100\%$. The False Positive Rate is given by $FPR = \frac{FP}{FP+TN} \times 100\%$. According the information given by Table I, the detection rate and the false positive rate are calculated and compared with GSAD [26] and PAYL. The results are shown in Table II.

TABLE II. PERFORMANCE COMPARISON OF VARIOUS IDSS

Systems	Number of Features	Detection Rate (DR%)	False Positive Rate (FPR%)
LDA-based IDS	300	100	3.7
GSAD	256^2	100	0.087
PAYL	256	98	0.1

The experimental results demonstrate that not all the features in MDM are essential for the detection of network attacks, and the observed packets can be classified using only a small number of significant features.

The proposed LDA-based feature selection approach is proven to be a good candidate for achieving the above mentioned task. It successfully transforms the original 256^2 dimensional feature domain to a relatively very low dimensional feature space while preserving the most significant information for the final detection. The features in the selected low dimensional feature space represent the most significant difference between the normal packets and the attack packets. By making use of the proposed LDA-based feature selection approach, we can not only significantly reduce the computational complexity of the detection process but also keep high detection rate.

V. CONCLUSIONS AND FUTURE WORK

This paper has proposed a LDA-based feature selection approach to reduce the computational cost of payload based anomaly IDS in attack detection. It is the first time that LDA is considered for payload-based feature selection. The approach not only extracts a set of low-dimensional features but also preserves most of the signification information for data classification

The proposed approach has been evaluated using DARPA 1999 IDS dataset and tested on HTTP traffic. It has achieved encouraging results with 100% detection rate and 3.7% false positive rate. However, the amount of selected significant features may grow to a large number when more types of attacks are considered. This is because more sets of significant features will be selected with respect to the increasing number of types of attacks. If there are not common features within the different sets of significant features, all of the features will be combined together and the number will multiply. Thus, we attempt to solve this issue by developing a fixed length identical significant feature set. Also, we will extend this research work to other attacks, e.g., Code-red worm, DDK and shell code attacks, as well as to other application-layer protocols, such as DNS, mail system etc.

REFERENCES

[1] CERT, "CERT Statistics", http://www.cert.org/stats/#notes, 2007.

[2] J. Kay, Low volume viruses: new tools for criminals, Network Security, 6, 2005, pp.16-18

[3] DE Denning, "An intrusion detection model", in: Proceedings of the 1986 IEEE Symposium on Security and Privacy (SSP 1986), IEEE Computer Society Press; pp. 118-133.

[4] TippingPoint: http://www.tippingpoint.com/

[5] V. Paxson, "Bro: a system for detecting network intruders in real-time", Computer Networks (Amsterdam, Netherlands: 1999), 31(23-24):2435–2463, 1999

[6] A. Patcha and J. M. Park, "An overview of anomaly detection techniques: existing solutions and latest technological trends", Computer networks, 2007.

[7] H. S. Javits and A. Valdes, "The NIDES statistical component: Description and justification", Technical report, SRI International, computer Science Laboratory, 1993.

[8] C. Taylor and J. Alves-foss, "NATE-Network Analysis of Anomaly Traffic Events, A Low-Cost approach", New Security Paradigms Workshop, 2001.

[9] J. Hoagland, SPADE, Silican Defence, http://www.silicondefence.com/software/spice, 2000.

[10] K. Wang and S. Stolfo, "Anomalous payload-based network intrusion detection", In Recent Advances in Intrusion Detection, RAID, pp. 203–222, September 2004.

[11] M. Mahoney and P. Chan, "Learning non stationary models of normal network traffic for detecting novel attacks", In Proc. SIGKDD 2002, pp. 376–385, 2002.

[12] M. Mahoney, "Network traffic anomaly detection based on packet bytes", In Proc. ACM-SAC, Melbourne FL, pp. 346– 350, 2003.

[13] D. Yang and H. Qi, "A network intrusion detection method using independent component analysis," 2008, pp. 1-4.

[14] H.C. Shih, et al., "Detection of Network Attack and Intrusion Using PCA-ICA," 2008, pp. 564-564.

[15] S. Singh and S. Silakari, "Generalized Discriminant Analysis algorithm for feature reduction in Cyber Attack Detection System," Arxiv preprint arXiv:0911.0787, 2009, pp. 173-180.

[16] V.A. Golovko, et al., "Dimensionality reduction and attack recognition using neural network approaches," 2007, pp. 2734-2739.

[17] V. Venkatchalam and S. Selvan, "Performance comparison of intrusion detection system classifier using various feature reduction techniques," International journal of simulation, vol. 9, no. 1, 2008, pp. 30-39.

[18] Y. Chen, et al., "Survey and taxonomy of feature selection algorithms in intrusion detection system," Springer, 2006, pp. 153-167

[19] C. Krügel, et al., "Service specific anomaly detection for network intrusion detection," Proc. Proceedings of the 2002 ACM symposium on Applied computing, ACM, 2002, pp. 201 - 208..

[20] K. Wang, et al., "Anagram: A content anomaly detector resistant to mimicry attack," Springer, 2006, pp. 226-248.

[21] N. Nwanze and D. Summerville, "Detection of anomalous network packets using lightweight stateless payload inspection," 2008, pp. 911-918.

[22] A. Jamdagni, et al., "Intrusion Detection Using Geometrical Structure," Proc. Frontier of Computer Science and Technology, 2009. FCST '09. Fourth International Conference on, 2009, pp. 327-333.

[23] A. Webb, Statistical pattern recognition, A Hodder Arnold Publication, 1999.

[24] Y. Chen, et al., "Pixel Structure Based on Hausdorff Distance for Human Detection in Outdoor Environments," Proc. Digital Image Computing Techniques and Applications, 9th Biennial Conference of the Australian Pattern Recognition Society on, 2007, pp. 67-72

[25] http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/data/index.html

[26] A. Jamdagni, et al., "Intrusion detection using GSAD model for HTTP traffic on web services," Proc. Proceedings of the 6th International Wireless Communications and Mobile Computing Conference, ACM, 2010, pp. 1193-1197.