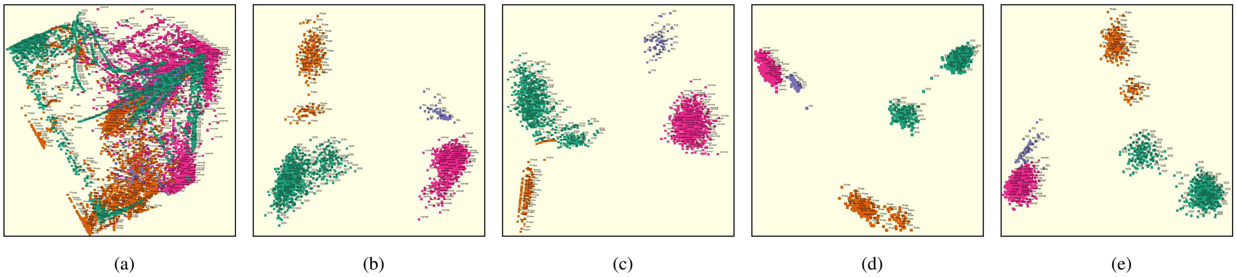
Applying Data Transformation to Derive Insights for Network Intrusion Detection

Dong Hyun Jeong
djeong@udc.edu
University of the District of Columbia

Soo-Yeon Ji
sji@bowiestate.edu
Bowie State University



(a) (b) (c) (d) (e)

Figure 1: PCA projections of (a) original dataset and DWT feature dataset with (b) Daubechies, (c) Biorthogonal, (d) Discrete Meyer, and (e) Symlets. The network traffic data are mapped with different color attributes as DoS (green), Probe (brown), R2L (purple), and Normal (red).

Abstract

Since network traffic become more prevalent and complex, it is important to design new innovative approaches to detect network anomalous activities for protecting our computing infrastructures. In this study, we analyzed network traffic data by transforming network traffic data to derive insights. Specifically, extracting hidden underlying patterns from the data is performed by applying a signal processing technique with statistical validation. Also, a visualization tool is designed to support an interactive understanding of the data. The effectiveness of our approach is validated with a broadly known intrusion dataset (called NSL-KDD).

1 Introduction

To protect our computing infrastructures, designing efficient intrusion detection techniques is important to be considered. A traditionally known network intrusion detection system detects network attacks by analyzing network packets at the network layer by comparing them to known attack patterns. However, this approach has a limitation of detecting unknown (or new) attacks. Therefore, detecting threats or intrusions by computationally analyzing network traffic patterns without having previous knowledge is emphasized. For this reason, researchers have proposed numerous intrusion detection techniques. However, there is a major limitation in the approach as of having a high false alarm rate (i.e. high false positives). To reduce the high false alarm rate, our approach emphasizes an integration of data transformations. Discrete Wavelet Transform (DWT) is applied to produce a new transformed data. Then, a statistical validation is performed to determine significant features. A visualization tool is also designed to identify clearly distinctive attack clusters with supporting an interactive visual analysis.

2 Approach

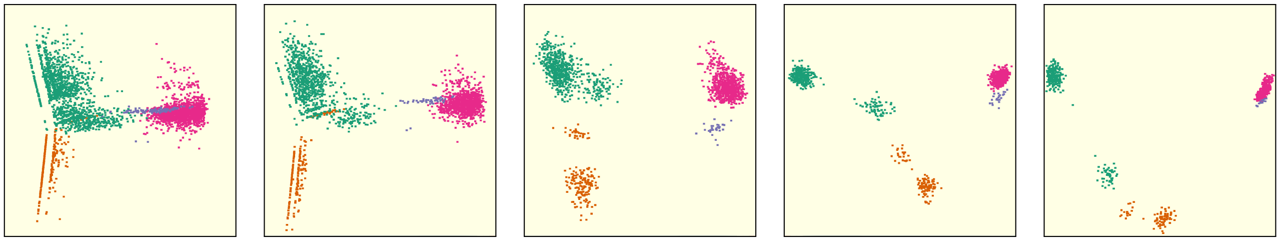
As we mentioned above, a publicly available intrusion detection dataset (called NSL-KDD) is used. The dataset contains 148,517 records - 125,973 (training set) and 22,544 (testing set) - with including 41 attributes (three nominal, six binary, and thirty-two numeric attributes). The dataset consists of normal activity and twenty-four attacks. These attacks are grouped into four major attack categories as DoS, R2L, U2R, and Probe. DoS attack represents any attempts to disable network access from remote machines (or computing resources). R2L indicates that a remote user gains access to local user accounts by sending packets to a computing machine over the network. Probe represents that network is scanned to gather information to find known vulnerabilities. U2R denotes that an attacker accesses normal users' accounts by exploring the system as an administrator. U2R is not considered in this study because the size of the data elements (119 records) is too small.

As the first data transformation procedure, a signal processing technique is used since it has a capability of discovering hidden patterns from the dataset. Specifically, DWT is applied to extract information from the dataset in different resolution levels. DWT uses two basis functions called wavelet function $\Psi(t)$ and scaling function $\varphi(t)$ to dilate and shift signals. The functions are applied to transform input data into a set of approximation coefficients and detail coefficients. There are various wavelet functions available as Daubechies, Coiflets, Symlets, Discrete Meyer, Biorthogonal, etc. Choosing an appropriate wavelet function that is closely matched to variations in the input dataset is considered as an important step. Also, determining an appropriate sliding window size (α), step (β), and wavelet level (γ) is critical when analyzing network traffic data [1, 2]. Often, researchers use time information to determine the size of the sliding window size. However, there is no optimal size for detecting network intrusions. To determine optimal values for α, β, and γ, we conducted an empirical study.

When applying DWT, different sizes of data records and attributes can be generated depending on applied wavelet functions and wavelet levels. For instance, when applying Daubechies 3 (i.e. db3) with $\alpha = 150$, $\beta = 50$, and $\gamma = 3$, total of 2,958 data records with 703 attributes are generated. 418 (59.5%) out of 703 attributes are selected as statistically significant attributes ($p < 0.01$). However, when using Discrete Meyer with $\alpha = 50$, $\beta = 20$, and $\gamma = 3$, total of 7,417 data records with 3,478 attributes are generated. 2,992 attributes (86%) are determined as significant features ($p < 0.01$). The determined DWT features are considered as dominant features that can be used to detect intrusions.

3 Visual Analysis

To represent the network traffic dataset, a visual analytics tool is designed. Since the dataset has too many attributes, it is difficult



(a) $\alpha = 50, \beta = 25, \gamma = 3$ (b) $\alpha = 100, \beta = 50, \gamma = 3$ (c) $\alpha = 150, \beta = 75, \gamma = 3$ (d) $\alpha = 200, \beta = 100, \gamma = 3$ (e) $\alpha = 300, \beta = 150, \gamma = 3$

Figure 2: An example empirical study result of identifying optimal values for sliding window size (α), step (β), and wavelet level (γ) with the wavelet function (i.e. Daubechies 3) to detect network intrusions clearly.

to represent them in a limited 2D display space. Therefore, PCA (Principal Component Analysis) computation is applied to reduce dimensions of the input features by identifying principal components. Figure 3 shows a screenshot of the designed tool. It consists of two views - (a) PCA projection view and (b) Data view. In the PCA projection view, computed PCA results of the network traffic dataset (or transformed dataset) are displayed along the first and second principal components (x and y axis, respectively). The Data view represents the original dataset in a parallel coordinates. To support an interactive analysis on the visually represented data elements, a set of interaction techniques (such as zooming, panning, and selection) are supported. Zooming and panning would be useful interaction techniques to understand the projected data elements. Selection is also supported to highlight the selected item(s) visible in two different views. Figure 3 shows an example when an R2L attack is selected in the Projection view. Its corresponding information is highlighted in the Data view. To support an interactive similarity measure, computationally identifying similar items is supported with providing four different similarity measurements as cosine similarity, Euclidean distance, extended Jaccard coefficient, and Pearson correlation coefficient. With the user selected data item(s), a similarity measurement is performed to find similar items. If any similar items are detected, they are highlighted with different colors. This feature is useful when analyzing anomalies appeared in different attack clusters.

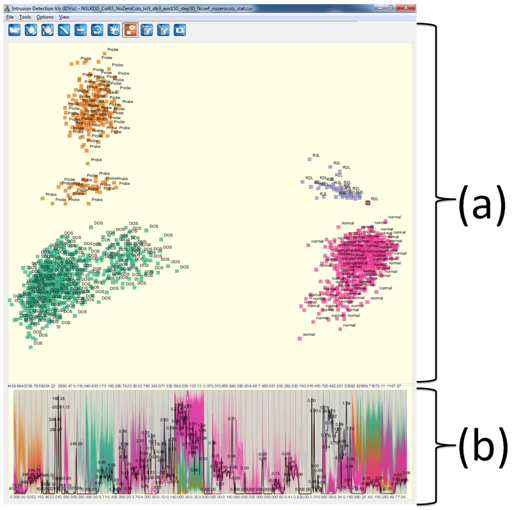


Figure 3: A screenshot of the designed intrusion detection analysis tool.

As shown in Figure 1a, it is difficult to identify a clear separation between normal and attacks. However, when looking at the projection of data transformed datasets (i.e. DWT features), a clear separation between them has appeared (see Figure 1b ~ 1e). From our empirical study (see Figure 2), we found that the attribute values ($\alpha = 150$, $\beta = 75$, and $\gamma = 3$) would be the optimal values for separating the attacks. We also found that when the sliding window size was small ($\alpha < 150$), the chance of false positive rate has increased. However, when the size was increasing ($\alpha >= 150$), a clear separation has emerged. From this study, we identified that a large sliding window often tends to remove unique characteristics of attacks. Therefore, Normal and R2L attack are often appeared in the same cluster (see Figure 2d and 2e). As shown in Figure 1b ~ 1e, these attribute values ($\alpha = 150$, $\beta = 75$, and $\gamma = 3$) were applied to determine the best wavelet function for network intrusion detection.

We identified that Daubechies 3 and 4 are the best-suited wavelet functions for detecting network intrusions since they generate well-separated clusters (see Figure 1b). Biorthogonal (Figure 1c) shows that some Probe attacks appeared in the DoS attack cluster. Discrete Meyer (Figure 1d) and Symlets (Figure 1e) are good for separating DoS and Probe attacks from R2L attack. However, The R2L attack is visible near to the normal network activity cluster (see the regions in left-top corner in Figure 1d and in left-bottom in Figure 1e). Interestingly, when applying Coiflets 1 function, we get a result that is similar to the one with Daubechies 3. It is because they use similar scaling functions.

4 Conclusion and Future Works

In this study, we emphasized the importance of applying data transformation. An interactive visual analytics tool is designed to help us understand the effectiveness of adapting data transformation. Since the tool provides a set of interaction techniques, an interactive visual analysis is possible to understand the represented data elements easily. Although our approach is good for network intrusion detection, a formal evaluation by measuring sensitivity and specificity should be performed. We also believe that our approach can be utilized to determine network intrusions in a real-time environment. To find an answer for this claim, we plan to test our approach with other known network traffic datasets.

Acknowledgements

This work was supported by the U.S. Army Research Office (ARO) - Grant No. W911NF-13-1-0143.

References

[1] P. Barford, J. Kline, D. Plonka, and A. Ron. A signal analysis of network traffic anomalies. In *Proceedings of the 2Nd ACM SIGCOMM Workshop on Internet Measurment*, IMW '02, pages 71–82, New York, NY, USA, 2002. ACM.

[2] C. T. Huang, S. Thareja, and Y. J. Shin. Wavelet-based real time detection of network traffic anomalies. In *Securecomm and Workshops, 2006*, pages 1–7, Aug 2006.