Feature-Based Transfer Learning for Network Security

Juan Zhao*, Sachin Shetty† and Jan Wei Pan‡
*Department of Electrical and Computer Engineering, Tennessee State University, Nashville, TN, USA
†Virginia Modeling, Analysis and Simulation Center, Old Dominion University, Norfolk, USA
‡Boeing Research and Technology, The Boeing Company, Huntsville, Alabama, USA
Email: jzhao1@tnstate.edu, sshetty@odu.edu, janwei.pan@boeing.com

***Abstract*—New and unseen network attacks pose a great threat to the signature-based detection systems. Consequently, machine learning-based approaches are designed to detect attacks, which rely on features extracted from network data. The problem is caused by different distribution of features in the training and testing datasets, which affects the performance of the learned models. Moreover, generating labeled datasets is very time-consuming and expensive, which undercuts the effectiveness of supervised learning approaches. In this paper, we propose using transfer learning to detect previously unseen attacks. The main idea is to learn the optimized representation to be invariant to the changes of attack behaviors from labeled training sets and non-labeled testing sets, which contain different types of attacks and feed the representation to a supervised classifier. To the best of our knowledge, this is the first effort to use a feature-based transfer learning technique to detect unseen variants of network attacks. Furthermore, this technique can be used with any common base classifier. We evaluated the technique on publicly available datasets, and the results demonstrate the effectiveness of transfer learning to detect new network attacks.**

***Index Terms*—Network attack detection, Machine learning, Transfer learning**

I. Introduction

Cyber-attacks are a growing concern in military and commercial networks due to the increased sophistication and variants of attacks, such as Denial of Service (DoS) tactics and zero-day threats. Conventional signature-based detection approaches are unable to address the increased variability of today's Cyber-attacks. Thus, there is a great need for developing an intelligent anomaly detection techniques to learn, adapt and detect threats in diverse network environments.

Machine learning techniques have been applied to improve the detection rate for malicious traffic based on establishing an explicit or implicit model that enables the patterns analyzed to be categorized [1][2][3]. However, they are still facing many challenges on detecting evolving attacks. Unsupervised anomaly detection used to detect new threats suffers from lower precision [4]. Data-driven supervised classifiers can achieve better precision, but they rely on the known malicious samples used in the training data. Once attacks change behaviors, the classifiers cannot work well and need to be retrained, as shown in Fig.1. With the continuously rising number of variants, this becomes a major bottleneck, as collecting sufficient datasets need great efforts. Moreover, when we need to incorporate new features from various network layers to tackle evolving attacks, we cannot directly apply the learned classifier on the testing data with a different feature space.

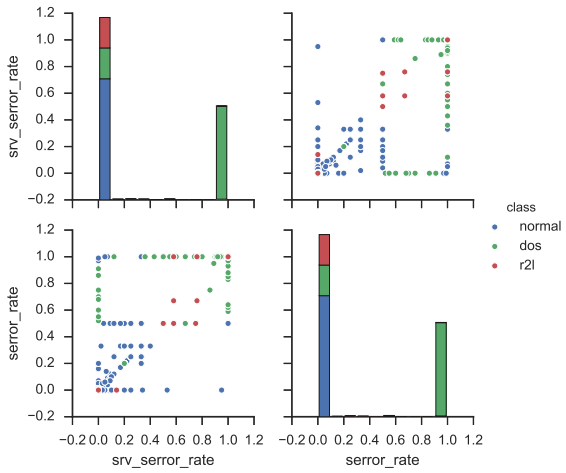


Fig. 1. Flow-based features are changing for different attacks. DoS (Denial of service) and R2L (Remote to local) exhibits different distributions in the network features. A decision boundary learned by distinguishing DoS from normal traffic can not be applicable in detecting R2L. **Features**: Serror_rate and srv_serror_rate describe the percent of connections with 'SYN' errors to the same host, and connections to the same destination port in the past two seconds respectively.

To address the problems, we propose using feature-based transfer learning techniques to enhance the adaptation and robustness of classifiers on detecting new threats. Transfer learning is a machine learning technique that can improve the prediction accuracy of the test (target) domain that has few or no labeled data by transferring the learned knowledge from a related training (source) domain [5]. The intuition behind transfer learning is inspired by the ability of human transitive inference and learning to extend what has been learned in one domain to a new similar domain [6]. Our study is motivated by two facts: 1) Most network attacks are variants of known network attack families that share similar traits [7]. Fig.2 illustrates an example of common traits, which is that network attacks need to scan a network of computers resulting in a smaller number of connections to the same destination compared with the connections in normal traffic, and 2) As networking protocols are standard, common features

978-1-5386-0595-0/17/$31.00 ©2017 IEEE

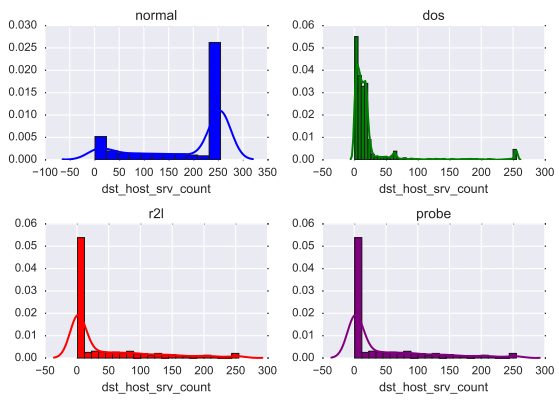


Fig. 2. DoS, R2L and Probe share similar distribution on feature dst_host_srv_count, which is the sum of connections to the same destination.

can be extracted for detecting network attacks. These common attributes viewed as a common latent structure suggests a good fit for applying transfer learning techniques.

In this paper, we present a transfer learning-enabled detection framework to detect previously unseen variants of attacks based on the learned attacks. We propose a feature-based heterogeneous transfer learning approach, called HeTL, which is an improvement on HeMap [8] to implement the framework. The main idea is to find optimized feature representations from both train and test network data. These representations can be achieved via the spectral transformation onto a common latent space where the difference of distributions can be reduced and the dimensions of the feature space can be equal. The new optimized feature representations are then fed to supervised learning algorithms such as decision trees, Naive Bayes, KNN, and SVM to train a more robust classifier. We will show that the classier trained on malicious samples from one category can successfully detect new samples from a different category. This way, the knowledge of the attack behavior is correctly transferred to the new domain. Compared to the baseline supervised classication without transfer learning, the proposed approach shows considerable improvements in classifying new types of network threats that were not part of the training data.

In short, the main contributions of our work are as follows:

1) We present a framework for detecting new unseen attacks by applying transfer learning techniques based on the known attacks. Using transfer learning is an important contribution on its own to enhance the adaptability of detection models.
2) To the best of our knowledge, we are the first effort to propose and evaluate feature-based transfer learning approaches for detecting new network attacks.
3) We evaluate the proposed technique on a benchmark network dataset NSL-KDD [9]. We compared our techniques with several baseline approaches. The experimental results show that the proposed technique achieves much better performance than any other baseline approach in detecting new attacks.

The rest of this paper is organized as follows: Section II reviews the related work. Section III presents the transfer learning-enabled detection framework and Section IV details the proposed transfer learning approach. We demonstrate our experimental settings and results in Section V. Finally, we conclude the work in Section VI.

II. Related work

One of the well-known techniques for network attack detection is the signature-based detection [10], which is based on an extensive knowledge of the particular characteristics of each attack, referred to as its signature. Another type of technique for network attack detection is the supervised learning-based technique [1][11]. Both of the study suffers from lower precision on detecting the new attacks since they mainly rely on the known examples of attacks. Transfer learning has recently gained attention due to several benefits: less effort in learning a new task and generation of robust learned models. Transfer learning has many successful applications in natural language processing and visual recognition and it is viewed as the promising area of machine learning. Bekerman et al. [2] mentioned transfer learning could improve the robustness in detecting unknown malware between non-similar environment. However, they did not present much detailed and formal work on this idea. The study in [12] applied an instance-based transfer learning approach in network intrusion detection. However, they required plenty of labeled data from target domain. Gao et al. [13] proposed a model-based transfer learning approach and applied it to the KDD99 cup network dataset. Both of these instance-based and model-based transfer learning approaches, unlike the feature-based approach, however, are heavily depending on the assumption of the homogeneous feature set. This is often not applicable to the network attack detection that typically exhibits heterogeneous features. Another advantage of feature-based approaches is its flexibility of adopting different base classifiers according to different cases, which motivated us to focus on deriving a feature-based transfer learning approach for the network attack detection study. To our best knowledge, this paper is the first effort to apply feature-based transfer learning approach for improving the robustness of network attack detection.

III. Overview of the Transfer Learning Framework

In this section, we present a transfer learning-enabled network attack detection framework (Fig.3) to enhance detecting new network attacks by transferring the knowledge learned from known network attacks. We use terms source domain and target domain, to denote the training and testing dataset in a machine learning task respectively. Both source domain and target domain data consist of normal as well as anomalous traffic records. We assume that the attack in the source domain is already known and labeled, and attacks in the target domain are new and unlabeled. The goal of the transfer learning framework is to use the source domain data to help differentiate new attacks from the target domain.

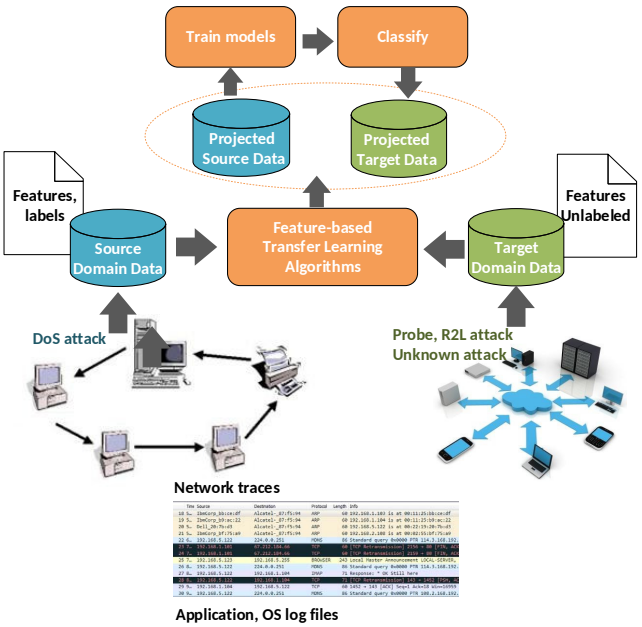


Fig. 3. Framework of proposed transfer learning-enabled network attack detection

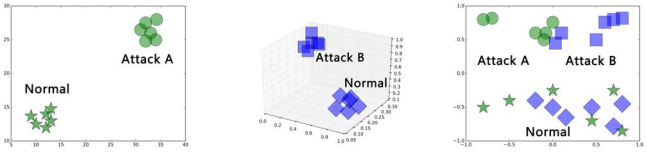
From a practical standpoint, the source and target domains can represent different or same network environments which are subject to different attacks that are captured at different times and at separate instances. In this paper, we primarily consider the latter scenario, wherein the source and target domain represent the same network but have different attacks. Unlike prior efforts [12][13], which assume that the source and target domains should have the same feature sets, our framework supports introducing new features in the target domain. This is relevant to evolving network attacks where the adversary may change the behavior, which results in identifying new features in the network or system layers. Thus, the source and target domains can have different attack distributions or feature sets.

The transfer learning framework consists of three main stages: (1) Extracting features from raw network data, (2) Learning representations with feature-based transfer learning, and (3) Supervised classification. In the first stage, features are extracted from the raw network trace data with statistic calculation of the network flow. Second, a good new feature representation is learned from both source and target domain via the feature-based transfer learning algorithms. The new representation will be feed to a common base classifier. The choice of a common base classifier here can be decision trees, random forest trees, SVM, KNN or Naive Bayes. The classifier trained on known samples of one category can detect the new samples from another different category.

IV. Transfer Learning Approach via Spectral Transformation

Problem formulation: We model the network attack detection as a binary classification problem, which is to classify each network connection as a malicious or as a normal connection. Suppose we are provided with source domain training examples $S = \{\vec{x_i}\}, \vec{x} \in \mathbb{R}^m$ that have labels $L_S = \{y_i\}$, and target domain data $T = \{\vec{u_i}\}, \vec{u} \in \mathbb{R}^n$, where both $\vec{x}$ and $\vec{u}$ are the feature extracted from network traffic. $\vec{x}$ and $\vec{u}$ are drawn from different distributions, $P_S(X) \neq P_T(X)$, and the dimensions of $\vec{x}$ and $\vec{u}$ are different, $\mathbb{R}^m \neq \mathbb{R}^n$. Our goal is to accurately predict the labels on T.

Note that our source data S and target data T have (1) different feature space $\mathbb{R}^m \neq \mathbb{R}^n$, and (2) different distribution $P_S(X,Y) \neq P_T(X,Y)$. Our approach is to explore a new common latent space by spectral transformation, where the distributions of malicious instances are similar and discriminative examples are still far apart. Our final goal is to learn the new representation of original source and target data in a $k-$dimensional latent semantic space, i.e.,$V_S \in \mathbb{R}^k, V_T \in \mathbb{R}^k$, so that we can classify malicious activities better using V_S, V_T instead of original S, T. The main idea is shown in Fig.4, as in the new projected common latent space (Fig.4(c)), the distributions of attack A and attack B are similar even though they look different in their respective original 2-dimension and 3-dimension spaces.



(a) Source domain (b) Target domain (c) Projected space

Fig. 4. Illustration of proposed feature space transformation concept.

We now discuss how to find the common latent subspace. The optimal subspace is defined as follows.

Optimization: Given the source data S and the target data T, we are looking for an optimal projection of S and T onto the optimal subspace V_S and V_T according to the following optimization objective:

$$\min_{V_S, V_T} \ell(V_S, S) + \ell(V_T, T) + \beta \cdot D(V_S, V_T), \quad (1)$$

where $\ell(*,*)$ is a distortion function that evaluates the difference between the original data and projected data. $D(V_S, V_T)$ denotes the difference between the projected data of the source and target domain. β is a trade-off parameter which controls the similarity between the two datasets.

Thus, the first two elements of (1) ensures the projected data preserve the structures of the original data as much as possible. We define $\ell(*,*)$ as follows:

$$\ell(V_S, S) = \|S - V_S P_S\|^2, \ell(V_T, T) = \|T - V_T P_T\|^2, \quad (2)$$

where V_S and V_T are achieved by a linear transformations with linear mapping matrices denoted as $P_S \in \mathbb{R}^{k \times m}$ and $P_T \in \mathbb{R}^{k \times n}$ to the source and target respectively. $\|X\|^2$ is the Frobenius norm which can also be expressed as matrix trace norm. In a different view, $P_S^T \in \mathbb{R}^{m \times k}$ and

$P_T^T \in \mathbb{R}^{n\times k}$ project the original data S and T into a k-dimensional space where the projected data are comparable ($\ell(V_S, S) = \|SP_S{}^T - V_S\|^2$). But in this case, this will lead to a trivial solution $P_S = 0, V_S = 0$. We thus apply (2). It can be viewed as a Matrix Factorization, which is widely known as an effective tool to extract latent subspaces while preserving the original datas structures.

We define $D(V_S, V_T)$ in terms of $l(*,*)$ as:

$$D(V_S, V_T) = \ell(V_S, V_T), \tag{3}$$

which is the difference between the projected target data and the projected source data. Hence, the projected source and target data are constrained to be similar by minimizing the difference function (3) .

Optimization

Substituting (2) and (3) into (1), we obtain the following optimization objective to minimize w.r.t V_S, V_T, P_S and P_T as follows:

$$\begin{aligned} &\min_{V_S^T V_S = I, V_T^T V_T = I} G(V_S, V_T, P_S, P_T) \\ =&\min_{V_S^T V_S = I, V_T^T V_T = I} \|S - V_S P_S\|^2 + \|T - V_T P_T\|^2 \\ &+ \beta \cdot (\|V_T - V_S\|^2). \end{aligned} \tag{4}$$

Obviously this is a not convex problem. Here we use gradient method to get the global minimums by iteratively fixing three of the matrices to solve the remaining one until convergence. For example, for solving V_T, fix V_T, P_T, P_S. Take the derivative of V_T, we can have the updates for V_T:

$$\frac{\partial G}{\partial V_T} = 2(V_T P_T P_T^T - T P_T^T + \beta(V_T - V_S)) \tag{5}$$

Similarly, for solving V_S, fix V_T, P_T and P_S. The corresponding gradients are given by:

$$\frac{\partial G}{\partial V_S} = 2(V_S P_S P_S^T - S P_S^T + \beta(V_S - V_T)) \tag{6}$$

For solving P_T, fix V_T, V_S and P_S. The corresponding gradients are given by:

$$\frac{\partial G}{\partial P_T} = (V_T^T V_T)^{-1} V_T^T T \tag{7}$$

For solving P_S, fix V_T, V_S and P_T. The corresponding gradients are given by:

$$\frac{\partial G}{\partial P_S} = (V_S^T V_S)^{-1} V_S^T S \tag{8}$$

The full algorithm is listed in Algorithm 1.

V. Performance Evaluation

In this section, we evaluate the effectiveness of transfer learning approach on a benchmark network attack dataset.

Algorithm 1: Heterogeneous Transfer Learning via Spectral Transformation

Input: T, S, β, k, learning rates α, steps = 1000
Output: V_T, V_S

Normalize $\mathbf{T}, \mathbf{S}$ **Initialize:** $V_T, V_S, P_T, P_S, step$
while *Optimized Function 4 not converge or step* $<$ *steps* **do**
  Update V_T by gradient descent with Eq. (5), $V_T = V_T - \alpha\frac{\partial G}{\partial V_T}$
  Update V_S by gradient descent with Eq. (6), $V_S = V_S - \alpha\frac{\partial G}{\partial V_S}$
  Update P_T by gradient descent with Eq. (7), $P_T = P_T - \alpha\frac{\partial G}{\partial P_T}$
  Update P_S by gradient descent with Eq. (8), $P_S = P_S - \alpha\frac{\partial G}{\partial P_S}$
  step++
end

TABLE I
Number of Instances in NSL-KDD

Class	Instances	Percentage
Normal	67343	53.46%
DoS	45927	36.46%
R2L	995	0.79%
Probe	11656	9.25%
U2R	52	0.04%

A. Dataset

We used NSL-KDD benchmark dataset [9], which contains extracted features from a series of TCP connection records on a local area network. Each sample in the dataset corresponds to a connection, which is labeled as either normal or an attack type. The dataset has 22 different attacks, which can be grouped by four main classes: Denial of service (DoS), Probe, Remote to local (R2L), and User to root (U2R). Table I provides the details of the attacks and their distribution in the training dataset. Since the portion of U2R is very small, we only focus on DoS, R2L, and Probe.

NSL-KDD contains 41 network features spitted into three groups: (1) basic features deduced from TCP/IP connections packet headers; (2) traffic features, usually extracted by flowing analyzing tools; and (3) content features, requiring processing of packet content.

B. Experimental setting

We model network attack detection as a binary classification problem to distinguish malicious traffic from normal traffic. To evaluate the transfer learning approach, we generate the dataset for source and target domain in different settings.

1) Detection of unseen new network attacks: To evaluate the performance of our transfer learning approach for detecting the unseen variants of attacks, we treat the problem as detecting attacks which are only present in a target domain network, but unseen in a source domain network. We assume the presence of one attack in the source domain and a different

attack in the target domain. Thus the distributions of attacks' feature values between source domain and target domain are different. Based on NSL-KDD, we have recreated three datasets, each of which contains a set of randomly selected normal examples and a set of attacks from one class. We set one of the datasets as the target domain and the other as the source domain. Thus, we mainly have three detection tasks: DoS→R2L (i.e. source domain DoS for training, target domain R2L for testing), DoS→Probe and Probe→R2L. We assume the feature spaces between source and target domain are the same. Table II and Table III show the performance comparison between the proposed transfer learning detection technique with HeTL and the baseline approaches. Fig.5 shows the ROC Curve. Fig.8 compares HeTL with other feature-based transfer learning approaches.

2) Network attacks with different feature spaces: The focus of this setting is to evaluate the performance of HeTL in detecting attacks with different feature spaces. The only difference from the previous setting is that the feature spaces for source and target domains are different. We use information gain to select the most relative features for the source and target domains, which results in different feature spaces for source and target domains. Traditionally, machine learning approaches are not effective in dealing with heterogeneous feature space. Therefore, we use manually mapping approach to transform the target data into the source feature space and evaluate its performance by using the classifiers, as the baseline approaches. Fig.6 shows the comparison of the transfer learning approach with baselines on DoS→R2L.

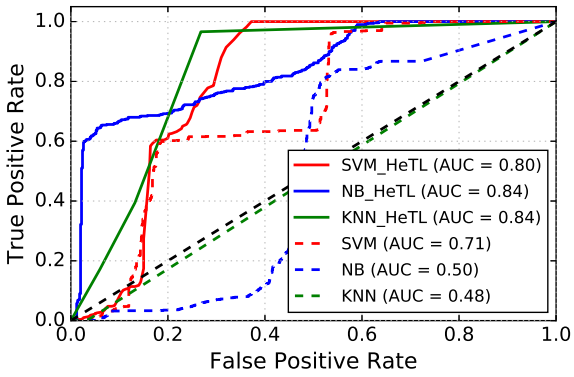


Fig. 5. ROC Curve on DoS→R2L

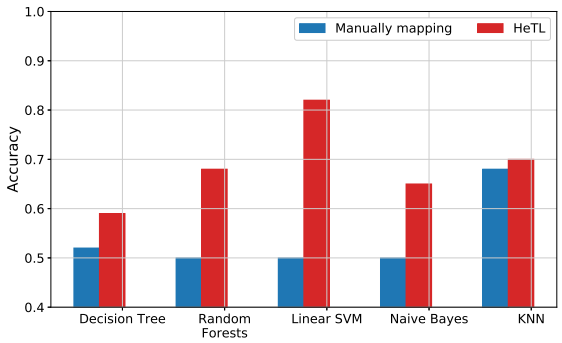


Fig. 6. Comparison on heterogeneous spaces on DoS→R2L

TABLE II
ACCURACY OF UNSEEN NETWORK ATTACK DETECTION

Datasets	Approach	CART	Random Forests	SVM	Naive Bayes	KNN
DoS→R2L	No_TL	0.53	0.50	0.49	0.36	0.49
	HeTL	0.81	0.78	0.81	0.72	0.78
DoS→Probe	No_TL	0.63	0.63	0.74	0.54	0.73
	HeTL	0.78	0.78	0.85	0.76	0.78
Probe→R2L	No_TL	0.52	0.50	0.47	0.65	0.51
	HeTL	0.73	0.75	0.75	0.71	0.78

TABLE III
F1 SCORE OF UNSEEN NETWORK ATTACK DETECTION

Datasets	Approach	CART	Random Forests	SVM	Naive Bayes	KNN
DoS→R2L	No_TL	0.12	0.0	0.0	0.0	0.0
	HeTL	0.83	0.80	**0.83**	0.74	0.81
DoS→Probe	No_TL	0.26	0.45	0.65	0.54	0.68
	HeTL	0.76	0.74	**0.84**	0.73	0.75
Probe→R2L	No_TL	0.12	0.01	0.04	0.57	0.08
	HeTL	0.67	0.77	0.79	0.77	0.78

C. Evaluation

1) Transfer learning vs non-transfer learning: For the first experimental setting, which aims to detect the new and unseen attacks, we compare the performance of the proposed transfer learning detection technique HeTL with common base classifiers without transfer learning approaches on three detection tasks. We have chosen decision tree (CART), random forests, linear SVM, Naive Bayes and KNN as the common base classifiers. From the results in Table II and Table III, we can see that the HeTL has significantly improved the performance in all cases. Compared with the baselines, HeTL improved accuracy of 35% - 50% and achieved much higher improvements in F1 -scores. From the ROC Curve as shown in Fig.5, we can see that HeTL improved the detection rates against the baselines. For the second experimental setting, Fig.6 shows the potential benefit of taking advantage of the heterogeneous feature space.

The results show that HeTL has a steady performance in the five common base classifiers, which means HeTL can work with various supervised learning classifiers. Moreover, We plot the distribution of the new feature representations in Fig.7 to demonstrate why HeTL can improve the peformance. We can see that HeTL successfully find a subspace that can make the distributions of different attacks similar, while still apart far from the normal behaviors, which helps finding the decision boundary to distinguish malicious from the normal.

2) Comparison of feature-based transfer learning approaches: We have evaluated HeTL with other feature-based transfer learning approaches, which are HeMap [8] and CORrelation Alignment (CORAL) [14] on network attack detection

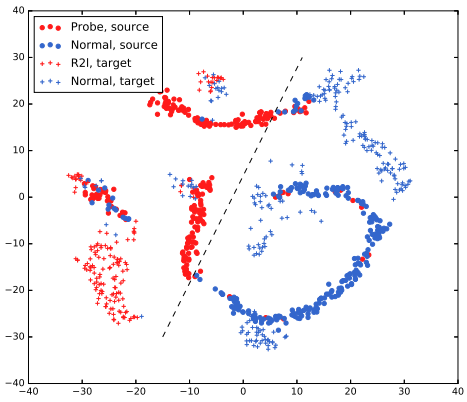


Fig. 7. Projected data on Probe→R2L. The source data keeps the original data structure and becomes similar to the target data. They share similar decision boundary in the projected space.

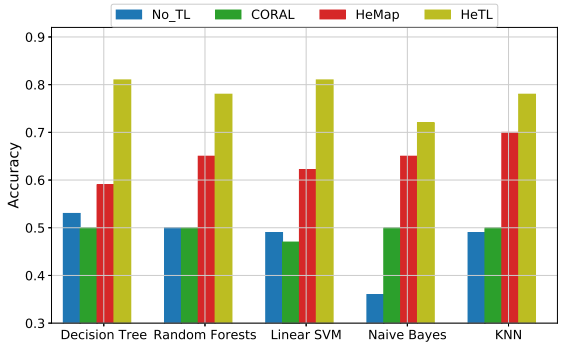


Fig. 8. Performance comparison of feature-based transfer learning on DoS→R2L

tasks. From results shown in Fig.8, we can see that HeTL has outperformed other feature-based approaches in all of the classifiers in the network attack detection task.

3) Parameters tuning: There are two tunable parameters: the similarity confidence parameter β and dimensions of the new feature space k. They can be manually set by empirical study or automatically set. There are several ways to automatically determine the optimum parameters. For example, the similarity confidence parameter β can be determined by computing similarity score between the source and target data. In this work, we use small labeled dataset (500 labeled) in the test domain to help find the optimum parameters. Future work includes a parametric study based on β and k.

VI. Conclusion

In this paper, we introduce a feature-based transfer learning framework and an approach called HeTL, to overcome the variants of attack causing the drop of detection performance. We evaluated the transfer learning approach on common base classifiers. The performance evaluation results show transfer learning approach improves the performance of detecting unseen new network attacks compared with baselines, and demonstrate that HeTL can support different feature space. The performance depends on the similarity confidence parameter β and dimensions of the new feature space k. For future work, we plan to develop analytical frameworks to derive the two parameters.

VII. Acknowledgments

This work was supported by Office of the Assistant Secretary of Defense for Research and Engineering (OASD (R& E)) agreement FA8750-15-2-0120 and Boeing Data Analytics agreement BRT-L1015-0006.

References

[1] R. Perdisci, W. Lee, and Feamster, "Behavioral clustering of HTTP-based malware and signature generation using malicious network traces," in *Prof. of the 7th USENIX conference on Networked systems design and implementation*, 2010, pp. 26–26.

[2] D. Bekerman, B. Shapira, L. Rokach *et al.*, "Unknown malware detection using network traffic classification," in *Communications and Network Security (CNS), 2015 IEEE Conference on*, Sep 2015, pp. 134–142.

[3] P. Garca-Teodoro, J. Daz-Verdejo, G. Maci-Fernndez, and E. Vzquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," *Computers and Security*, vol. 28, no. 12, pp. 18 – 28, 2009.

[4] K. Bartos, M. Sofka, and V. Franc, "Optimized invariant representation of network traffic for detecting unseen malware variants," in *USENIX Security 2016*. Austin, TX: USENIX Association, 2016, pp. 807–822.

[5] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[6] K. D. Feuz and D. J. Cook, "Transfer Learning Across Feature-Rich Heterogeneous Feature Spaces via Feature-Space Remapping (FSR)," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 1, pp. 3:1–3:27, 2015.

[7] D. Lin, "Network Intrusion Detection and Mitigation against Denial of Service Attack," *WPE-II Written Report*, no. January, pp. 1–28, 2013.

[8] X. Shi, Q. Liu, W. Fan *et al.*, "Transfer learning on heterogeneous feature spaces via spectral transformation," in *2010 IEEE International Conference on Data Mining*, Dec 2010, pp. 1049–1054.

[9] NSL-KDD, "UNB IUNB ISCX NSL-KDD DataSet," 2016. [Online]. Available: http://www.unb.ca/research/iscx/dataset/iscx-NSL-KDD-dataset.html

[10] H. Han, X.-L. Lu, and L.-Y. Ren, "Using data mining to discover signatures in network-based intrusion detection," in *Machine Learning and Cybernetics, 2002. Prof.. 2002 International Conference on*, vol. 1, 2002, pp. 13–17 vol.1.

[11] A. Valdes, A. Valdes, K. Skinner *et al.*, *Recent Advances in Intrusion Detection*, 2000, vol. 1907.

[12] S. Gou, Y. Wang, L. Jiao *et al.*, "Distributed Transfer Network Learning Based Intrusion Detection," in *2009 IEEE International Symposium on Parallel and Distributed Processing with Applications*, 2009, pp. 511–515.

[13] J. Gao, W. Fan, J. Jiang *et al.*, "Knowledge transfer via multiple model local structure mapping," in *Prof. of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '08. New York, NY, USA: ACM, 2008, pp. 283–291.

[14] B. Sun, J. Feng, and K. Saenko, "Return of Frustratingly Easy Domain Adaptation," in *AAAI'16*, 2016.