2015 IEEE Trustcom/BigDataSE/ISPA

Unsupervised Feature Selection Method for Intrusion Detection System

Mohammed A. Ambusaidi, Xiangjian He* and Priyadarsi Nanda
School of Computing and Communications, Faculty of Engineering and IT, University of Technology, Sydney, Australia
Email: Mohammed.A.AmbuSaidi@student.uts.edu.au, {Xiangjian.He, Priyadarsi.Nanda} @uts.edu.au

***Abstract*—This paper considers the feature selection problem for data classification in the absence of data labels. It first proposes an unsupervised feature selection algorithm, which is an enhancement over the Laplacian score method, named an Extended Laplacian score, EL in short. Specifically, two main phases are involved in EL to complete the selection procedures. In the first phase, the Laplacian score algorithm is applied to select the features that have the best locality preserving power. In the second phase, EL proposes a Redundancy Penalization (RP) technique based on mutual information to eliminate the redundancy among the selected features. This technique is an enhancement over Battiti's MIFS. It does not require a user-defined parameter such as β to complete the selection processes of the candidate feature set as it is required in MIFS. After tackling the feature selection problem, the final selected subset is then used to build an Intrusion Detection System. The effectiveness and the feasibility of the proposed detection system are evaluated using three well-known intrusion detection datasets: KDD Cup 99, NSL-KDD and Kyoto 2006+ dataset. The evaluation results confirm that our feature selection approach performs better than the Laplacian score method in terms of classification accuracy.**

Keywords*—Supervised feature selection, Unsupervised feature selection, Mutual information, Intrusion detection system.

I. Introduction

Feature selection is a technique for eliminating irrelevant and redundant features and selecting the most optimal subset of features that produce a better characterisation of patterns belonging to different classes. The feature selection problem has been around since the early 1970's. Due to its computational complexity, it still remains an open problem for researchers. Feature selection reduces computational cost, facilitates data understanding, improves the performance of modelling and prediction and speeds up the detection process [1].

A feature f_i in a feature space is relevant to the class if it embodies useful information about the class and its removal degrades the performance of the classification. The irrelevant feature is the one that does not contain any useful information about the class and its existence degrades the performance of the classification [2]. An irrelevant feature can be a redundant feature or a noisy feature. The redundant feature cannot provide any additional information to the classification after selecting the best subset of features because another feature has already given the same information. The noisy feature, which is not redundant does not contain any information about the class.

In accordance with the existence of label of data or not, feature selection techniques are generally classified into three groups: supervised, semi-supervised and unsupervised feature selection. Supervised and semi-supervised methods are usually applied on labeled data, while the unsupervised method is more appropriate for unlabeled data [3]. However, many real-world applications do not contain any label, hence, the unsupervised feature selection process becoming difficult and hard to achieve [4]. In this work the focus will be on unsupervised feature selection.

Several attempts have been made to develop an intelligent unsupervised feature selection technique which can utilise unlabeled data. The variance score method is one of the simplest unsupervised feature selection methods that calculates the variance of each of the features individually and selects the ones that have larger variance values [5]. Another unsupervised feature selection method is the Laplacian score [6]. Unlike the variance score algorithm, the Laplacian score not only selects the features with high variances, but also investigates the locality preserving power of every feature in the data. In many applications (such as many real-world applications), extracting the local structure information is very important in order to find the best features in the data [7], [6]. These methods, however, neglect the redundancy among selected features, so they select many redundant features, and affect the classification performance. This paper addresses this issue.

The key contribution of this paper is to develop an Extended version of the Laplacian score method, EL in short. EL proposes a redundancy penalization technique to eliminate redundancy among the selected features. This technique is an enhancement over Battiti's MIFS [8]. It does not require a user-defined parameter such as β to complete the selection processes of the candidate feature set as it is required in MIFS. After tackling the feature selection problem, the best selected subset of features is then used to train the classifier and build our intrusion detection system. Finally, we verify the effectiveness of the proposed detection system combined with EL by several experiments on three well known intrusion detection datasets: the KDD Cup 99, NSL-KDD and Kyoto 2006+ dataset. The experimental results of our method are compared using classification accuracy.

This paper is organized as follows: Section II reviews briefly the concept of mutual information and some related feature selection based on mutual information. Section III provides a description of the Laplacian score algorithm. Section IV discusses the proposed unsupervised feature selection method. Section V details our detection framework showing different detection stages involved in the proposed scheme. Section VI presents the experimental details and results. Fi-

978-1-4673-7952-6/15 $31.00 © 2015 IEEE
DOI 10.1109/Trustcom.2015.387

nally, a summary to the paper is drawn in Section VII.

II. Background on Mutual Information

The key concept of mutual information is from information theory which was proposed in 1948 by Shannon [9]. It describes the amount of information shared between two random variables. It is a symmetric measure of the relationship between two random variables, and it yields a non-negative value [10]. A zero value of MI indicates that the two observed variables are statistically independent. Given two random variables $X = \{x_1, x_2, ..., x_n\}$ and $Y = \{y_1, y_2, ..., y_n\}$, where n is the total number of samples, the mutual information between variables X and Y is defined as:

$$I(X;Y) = H(X) - H(X|Y) \tag{1}$$

where $H(X)$ is the uncertainty of X and $H(X|Y)$ is the conditional entropy, which are defined as

$$H(X) = -\sum_{x \in X} P_X(x) \log P_X(x) \tag{2}$$

$$H(X|Y) = -\sum_{x,y} P_{X,Y}(x,y) \log P_{X|Y}(x|y) \tag{3}$$

where $p(x)$ is the probability density function of X. To quantify the amount of knowledge on variable *X* provided by variable *Y* (and vice versa), mutual information can be defined as follows.

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{4}$$

where $p(x, y)$ is the joint probability density function of X and Y. From Equation (4), a high value of $I(X;Y)$ indicates both X and Y are closely related; otherwise, a zero value of $I(X;Y)$ means both X and Y are independent.

As stated above, a feature is relevant to the class if it contains important information about the class; otherwise it is irrelevant. Since mutual information is good at quantifying the amount of information shared between two random variables, it is often used to determine the relevance between features and the output class. Under this context, features with high predictive power are the ones that have larger mutual information $I(C; f)$. On the contrary, in the case of $I(C; f)$ equal to zero, the feature f and the Class C are proven to be independent from each other. This suggests that feature f contains redundant information.

Recently, mutual information has been used by a number of researchers to develop supervised feature selection methods [8], [11], [12], [13], [14]. Battiti in [8] harnessed MI between inputs and outputs for a single selection of features by calculating the $I(C; f_i)$ and $I(f_s, f_i)$, where f_s and f_i are candidate features and C is the class label. MIFS selects the feature that maximizes $I(C; f_i)$, which is the amount of information that feature f_i carries about the class C, and is corrected by subtracting a quantity proportional to the MI with the features selected previously. MIFS is a heuristic incremental search algorithm and the selection process continues until a desired number of R inputs are selected. Equation (5) shows the evaluation function of MIFS.

$$J_{MIFS} = I(C; f_i) - \beta \sum_{f_s \in S} I(f_i; f_s), \tag{5}$$

where β is a user-defined parameter that is apply to regulate the relative significance of the redundancy between the current feature and the set of previously selected features.

As can be seen, Equation (5) consists of two terms. The left-hand side term, $I(C; f_i)$, represents the amount of information that feature f_i carries about the class C. A relevant feature is the one that maximizes this term. The right-hand side term, $\beta \sum I(f_s; f_i)$, is used to eliminate the redundancy among the selected features.

In the follow-up research, various methods have been proposed to enhance Battiti's MIFS. Most of the studies have been conducted on the right-hand side term of Equation (5). Kwak and Choi in [11] made a better estimation of MI between input features and output classes and proposed a greedy selection algorithm named MIFS-U, in which U stands for uniform information distribution. MIFS-U shows a better estimation of $I(C; f_i)$ than MIFS. The algorithm of MIFS-U differs from that of MIFS in the right-hand side term as shown in Equation (6).

$$J_{MIFS-U} = I(C; f_i) - \beta \sum_{f_s \in S} \frac{I(C; f_s)}{H(f_s)} I(f_i; f_s) \tag{6}$$

Despite the redundancy parameter β used in the aforementioned methods to help to control the redundancy among features, it remains an open question on how to choose the most appropriate values for these parameters. If the chosen value is too small, the redundancy between input features is not taken into consideration and therefore both relevant and redundant features are involved in the selection processes. If the chosen value is too large, the algorithms only consider the relation between input features rather than the relation between each input feature and the class [14]. Thus, it is hard to determine the value of the parameter. In addition, both MIFS and MIFS-U neglect the influence of the number of selected features. This reduces the influence of $I(C; f_i)$ on Battiti's MIFS and Kwak's MIFS-U when the term on the right-hand side in in both methods increases, which is because this term is a cumulative sum [13]. This results in the irrelevant features being selected into the set S.

These limitations have been studied by Amiri in [15] and proposed Modified version of MIFS, MMIFS in short. MMIFS set the value of parameter β to be equal to $\beta^{'} / \mid S \mid$, where $\beta^{'}$ is the redundancy parameter, as shown in Equation (7).

$$J_{MMIFS} = I(C; f_i) - (\frac{\beta^{'}}{\mid S \mid}) \sum_{f_s \in S} I(f_i; f_s), \tag{7}$$

where $|S|$ is the cardinality of the set S, which is used to control the influence of the number of selected features since the right-hand side of the algorithm is a cumulative sum.

However, in the case of $\beta = \beta^{'}/\mid S \mid$ then MMIFS are equal to Battiti's MIFS. Therefore, the unbalance between the left and right hand sides in Equation (7) remains unsolved totally in MMIFS [15]. This might result in selecting irrelevant features. In addition, similar to Battiti's MIFS and Kwak's MIFS-U, selecting an appropriate value for the parameter $\beta^{'}$ in MMIFS remains an open question. In addition, all of these algorithms are supervised feature selection methods. These methods require labeled data. However, labeled data are not always available and also hard or expensive to obtain which makes these methods not applied to such data [4]. Therefore, in order to utilize unlabeled data, we propose an unsupervised feature selection method based on mutual information. This method removes the burden of setting an appropriate value for β and keeps the values of the right-hand side of our evaluation function within the range of [0,1]. This is helpful in practice since there is no specific guide on how to select the best value for this parameter. The proposed method is a modified version of Laplacian Score method, which ignores the redundancy among features. Next section introduces the Laplacian score algorithm.

III. Laplacian Score

To explain the Laplacian Score, we refer to the definition proposed in [6]. Laplacian Score (LS) is fundamentally based on Laplacian Eigenmaps [16] and Locality Preserving Projection [17]. The basic idea of LS is to evaluate the features according to their locality preserving power. In Section III, we re-state the algorithm to calculate the Laplacian Score as shown in [6].

The Algorithm: Let $x_p = [f_{1p}, f_{2p}, f_{3p}, \ldots, f_{np}]$, be the p-th traffic sample in this paper, where $p = 1, 2, \ldots, P$. Then, f_{ip} denotes the p-th sample of the i-th feature. Let L_i denote the Laplacian Score of the i-th feature, where $i = 1, ..., n$. The algorithm can be stated as follows.

1) Construct a nearest neighbor graph with P nodes. The p-th node is denoted by x_p. We put an edge between nodes p and q if x_p and x_q are "close", i.e. x_p is among k nearest neighbors of x_q or x_q is among k nearest neighbors of x_p. When the label information is available, one can put an edge between two nodes sharing the same label.
2) If nodes p and q are connected, put $S_{pq} = e^{-\frac{||x_p - x_q||^2}{t}}$, where t is a suitable constant. Otherwise, put $S_{pq} = 0$. The weight matrix S of the graph models the local structure of the data space.
3) For the i-th feature, we define: $\mathbf{f}_i = [f_{i1}, f_{i2}, ..., f_{iP}]^T$, $D = diag(S\mathbf{1})$, $\mathbf{1} = [1, \ldots, 1]^T$, $L = D - S$ where the matrix L is often called graph Laplacian [18]. Let

$$\check{\mathbf{f}}_i = \mathbf{f}_i - \frac{\mathbf{f}_i^T D\mathbf{1}}{\mathbf{1}^T D\mathbf{1}}\mathbf{1} \tag{8}$$

4) Compute the Laplacian Score of the i-th feature as follows.

$$L_i = \frac{\check{\mathbf{f}}_i^T L \check{\mathbf{f}}_i}{\check{\mathbf{f}}_i^T D \check{\mathbf{f}}_i} \tag{9}$$

IV. Modified Laplacian Score

To ensure the values of L_i and mutual information are not vary greatly, both values are adapted to the range [0,1]. Therefore, in this paper, a linear transformation normalisation to the value of L_i in Equation (9) is used as follows.

$$NL_i = \frac{L_i - L_{min}}{L_{max} - L_{min}} \tag{10}$$

where L_{min} and L_{max} are the minimum and maximum values of $\{L_1, L_2, ..., L_n\}$, respectively.

As discussed above, the Laplacian score does not take into consideration the redundancy among selected features. To address this issue, a scheme is proposed to eliminate redundancy among the selected features based on mutual information and appended to the Laplacian score.

Given a features set $F = \{f_1, f_2, \ldots, f_n\}$, where n is the total number of features, the task is to select the best subset of features $G = \{g_1, g_2, \ldots, g_{|G|}\}$, where $|G|$ is the number of selected features. The scheme is to normalise the value of mutual information between a candidate feature and the set of previously selected features by the entropies of the selected features as shown in Equation (11) in order to select the m-th feature, g_m, from $F \setminus \{g_1, g_2, ..., g_{m-1}\}$.

$$RPI(f_i; G) = \frac{1}{m-1}\sum_{j=1}^{m-1}\frac{I(f_i; g_j)}{H(g_j)}. \tag{11}$$

$$g(m) = \underset{f_i}{\operatorname{argmax}}(NL_i - RPI(f_i; G)), \tag{12}$$

where NL_i represents the normalised Laplacian score of the i feature as shown in Equation (10).

The overall procedure of EL algorithm is as follows.

Algorithm 1 Overall procedure of EL

Input: Feature set $F = \{f_i, i = 1, ..., n\}$, R : the number of selected features, $R \leq n$.
Output: G- the selected feature subset.
1. Initialization: set $G = \phi$.
2. Calculate NL_i $(i = 1, ..., n)$ according to Equation (9) and Equation (10) for each feature in F.
3. Select the feature f_i that maximises NL_i.
Set $F \longleftarrow F \setminus \{ f_i \}$; $G \longleftarrow G \cup \{ f_i \}$.
4. while $|G| < R$ **do**
  for *each feature* $f_i \in F$ **do**
    Calculate $RPI(f_i; G)$ in Equation (11) for all pairs of $(f_i; G)$.
  end
  Using Equation (12) select $g(m)$.
  Set $F \longleftarrow F \setminus \{ g(m) \}$ *and* $G \longleftarrow G \cup \{ g(m) \}$.
end
return G

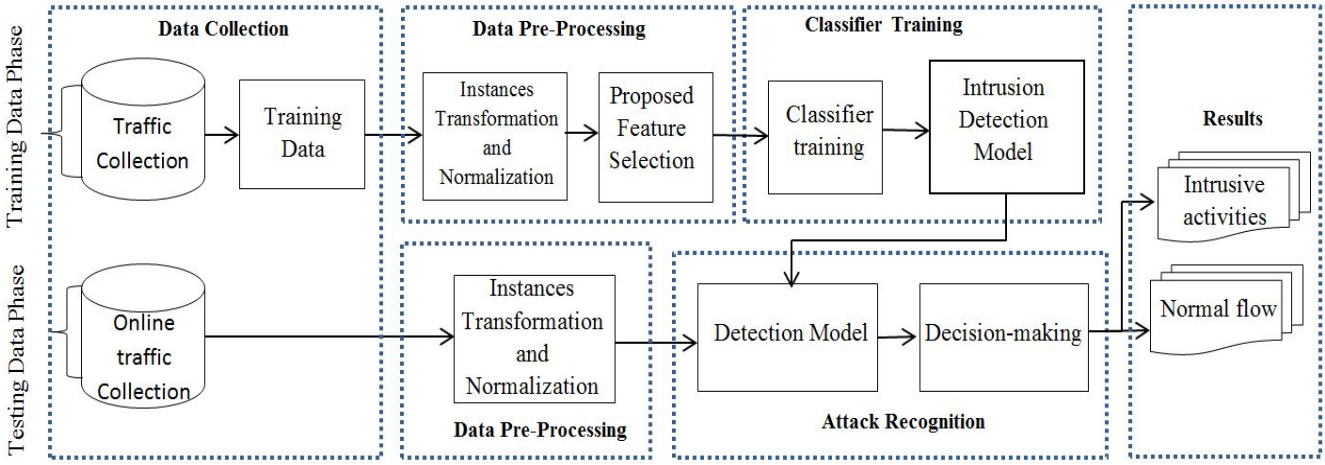

Fig. 1: The framework of the proposed intrusion detection system

V. Intrusion Detection Based on Unsupervised Feature Selection

The framework proposed in this chapter differs from the ones proposed in the previous chapters in the pre-selection stage, in which the proposed unsupervised feature selection is applied. The framework of the proposed detection model is shown in Figure 1. It can be seen from the figure that the detection framework is comprised of four main stages:

- Data Collection. It is the first and most important stage to intrusion detection where a sequence of network packets is collected.
- Data Pre-processing. In this stage, the obtained training and test data from the data collection stage are first pre-processed to generate basic features. This phase involves three main steps. The first step is data transferring, in which every symbolic feature in a dataset is first converted into a numerical value. The second step is data normalisation, in which each feature in the data is scaled into a well-proportioned range to eliminate the bias in favour of features with greater values from the dataset. The third step is feature selection, in which the proposed feature selection algorithm is used to nominate the most important features that are then used to train the classifier and build the intrusion detection model.
- Classifier Training. In this stage, the classifier is trained. Once the best subset of features is selected, this subset is then passed into the classifier training stage where a specific classification method is employed.
- Attack Recognition. In this stage, the trained model is used to detect intrusions on the test data. After completing all the iteration steps and the final classifier is trained which includes the most correlated and important features, the normal and intrusion traffics can be recognised by using the saved trained classifier. The test data is then taken through the trained model to detect attacks.

One can find more details about these stages in [19].

VI. Experiments and Results

To demonstrate the effectiveness of the proposed feature selection algorithm, three well-known intrusion detection datasets are used to assess and compare the performance of the proposed algorithm against the Laplacian score method. These datasets are the KDD Cup 99 datasets [20], NSL-KDD datasets [21] and Kyoto 2006+ datasets [22]. The results achieved by the proposed algorithm are compared with the results obtained by the Laplacian score and one of the existing unsupervised anomaly IDSs. To evaluate the effectiveness of the proposed algorithm, we perform binary classification and multi-classification. Two classifiers are used to serve the purpose of evaluations and comparisons, and they are the nearest neighborhood classifier (1NN) and Support Vector Machine (SVM) (LIBSVM package [23]). The details of the datasets are listed in Table I.

During the experiments, the value of R is given by the user in advance. To select the best value of k we have conducted several experiments and we set $k = 4$ for both the Laplacian Score and our proposed EL algorithm.

TABLE I: Summary of Datasets used in our experiments

Dataset	# Sample	# feature	# Class
KDD Cup 99	100,000	41	5
NSL-KDD	100,000	41	2
Kyoto 2006+	100,000	23	2

A. Benchmark datasets

Currently, there are only a few number of public datasets for intrusion detection evaluation. Therefore, we select the aforementioned datasets for our experiments since all of them are frequently used in literatures.

The KDD Cup 99 dataset is one of the most popular intrusion detection datasets that is widely applied to evaluate

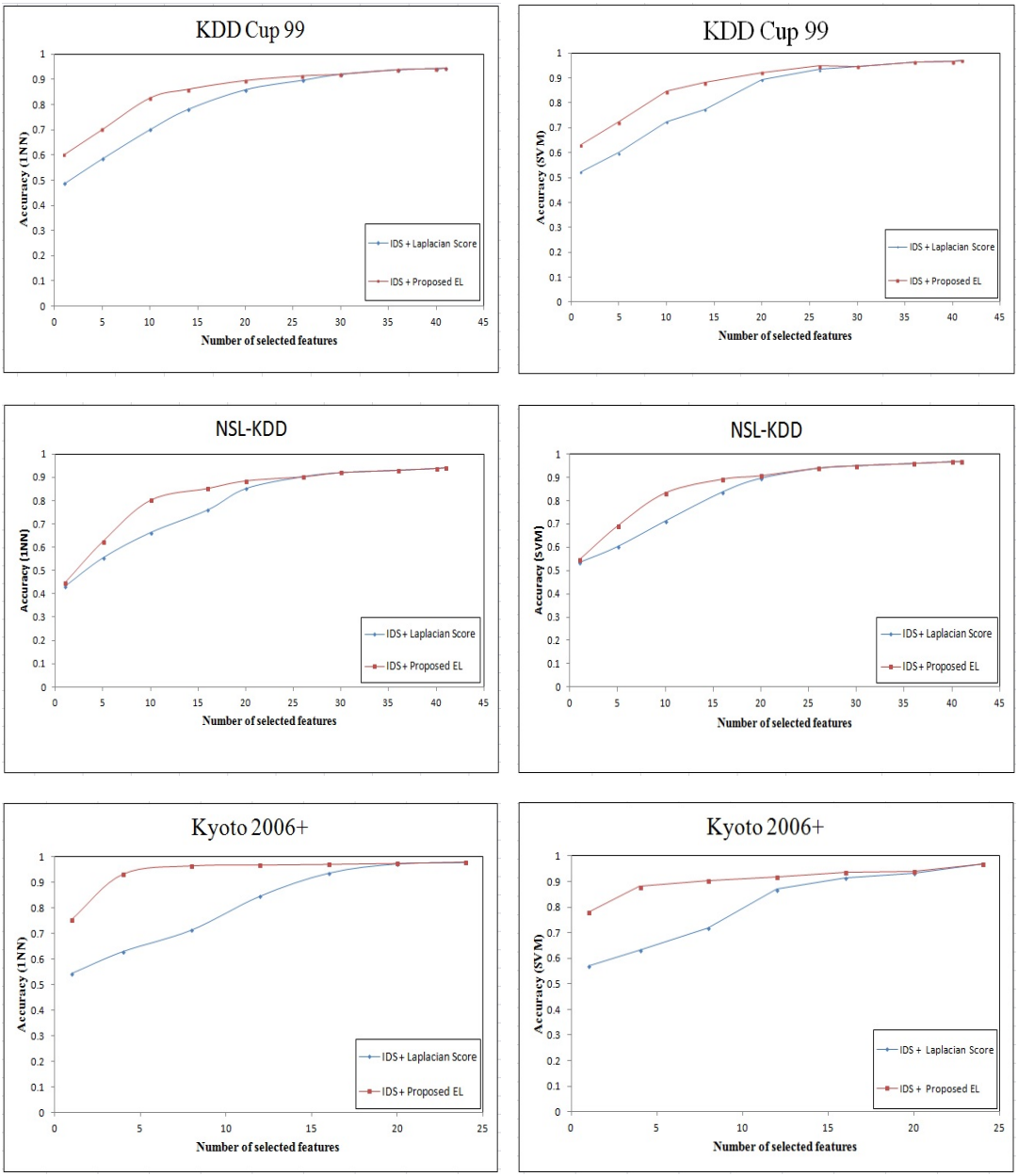

Fig. 2: Effect of number of selected features on IDS datasets with the two classifiers

the performance of IDSs [24]. It contains training data, "10% KDD Cup 99", with approximately five millions data connection records and test data, "kddcup testdata", with about two millions data connection records. KDD Cup 99 consists of five different classes, normal and four types of attack (i.e., DoS, Probe, U2R, R2L).

The NSL-KDD is a new revised version of KDD Cup 99 that has been proposed by Tavallaee et al. in [21]. This dataset addresses some to the problems included in the KDD Cup 99 dataset such as the huge number of redundant records in KDD Cup 99 data. The training and test datasets of NSL-KDD dataset consist of approximately 125,973 and 22,544 connection records respectively. Similar to the KDD Cup 99 datasets, each record in these datasets has 41 different quantitative and qualitative features.

The Kyoto 2006+ dataset was presented by Song et al. [22]. The dataset covers over three years of real traffic data collected from both honeypots and regular servers that are deployed at Kyoto University. It consists of approximately 50,033,015 normal sessions, 43,043,255 attack sessions and 425,719 sessions were unknown attacks. Each connection in the dataset is unique with 23 features. For our experiments, we select samples form the data of the days 2009 August 27, 28, 29, 30 and 31, which contain the latest updated data.

TABLE II: Detection accuracy of the two IDSs based on EL and the Laplacian score algorithms using 1NN and SVM classifier

#R	1NN		SVM	
	IDS + Laplacian Score	IDS + Proposed EL	IDS + Laplacian Score	IDS + Proposed EL
KDD Cup 99 ($n = 41$)				
4	57.29 ± 2.39	**68.39 ± 2.19**	59.95 ± 4.73	**69.93 ± 3.49**
8	65.22 ± 1.95	**80.22 ± 1.68**	67.08 ± 3.56	**82.19 ± 3.04**
12	72.42 ± 1.42	**84.42 ± 1.04**	76.88 ± 3.09	**87.58 ± 2.33**
16	73.33 ± 0.89	**87.33 ± 0.59**	88.13 ± 1.51	**90.36 ± 1.17**
NSL-KDD ($n = 41$)				
4	50.19 ± 3.35	**59.29 ± 3.33**	56.73 ± 4.81	**67.19 ± 3.85**
8	60.35 ± 2.47	**79.13 ± 2.92**	65.99 ± 3.32	**79.79 ± 3.43**
12	71.14 ± 2.28	**83.09 ± 1.09**	74.17 ± 2.99	**86.17 ± 2.32**
16	75.95 ± 2.02	**85.19 ± 0.93**	83.90 ± 2.51	**89.35 ± 2.15**
Kyoto 2006+ ($n = 23$)				
2	59.35 ± 2.10	**90.38 ± 1.46**	60.18 ± 2.17	**87.12 ± 1.26**
4	62.91 ± 1.16	**93.12 ± 1.23**	63.21 ± 1.92	**88.03 ± 1.23**
6	64.99 ± 1.02	**94.66 ± 0.80**	65.10 ± 1.40	**89.12 ± 1.12**
8	71.33 ± 0.48	**96.38 ± 0.46**	72.03 ± 1.31	**90.46 ± 0.56**

For experimental purposes, we randomly select 100,000 samples from each dataset. In order to decrease the random selection effect, all the experimental results in this paper are the averages of 10 runs. To avoid the bias in favor of features with greater values in all datasets, every feature within each record is normalized by the respective maximum value and falls into the same range of [0,1].

B. Results and discussion

In order to investigate the performance of our proposed feature selection algorithm, we build two intrusion detection systems based on our EL and the Laplacian score method. The aim is to further examine the advantages of removing redundancies among the selected features. We conducted our experiments on the three IDS datasets and compare the results achieved by the two detection systems. The experimental results about classification accuracies on these datasets are presented in Figure 2 and Table II.

Figure 2 plots the classification accuracies of 1NN and SVM achieved using both EL and the Laplacian score with R increasing from 1 to n. The x axis represents the number of selected features and y axis represents the classification accuracy. The figure shows that, in general, the classification accuracy improves when the number of selected features increases. It can be seen from the figure that the curve of our proposed EL method is above the curve of the Laplacian score method in all three datasets for almost all R values. That is because EL takes into consideration the redundancies among features and thus can select features with smaller redundancies. Note that when $R = n$, both systems achieved almost the same accuracy.

Table II summarizes the average classification accuracies using four different values of R on each dataset. The table shows clearly that the results obtained using EL are better than those obtained from the Laplacian Score method on all datasets in most of the cases.

C. Additional Comparison

The performance of our detection model using EL method is further compared with an unsupervised anomaly IDS, varGDLF in short, proposed by Fan et al. in [25]. Based on our knowledge, there is a small effort has been done to develop IDS that can utilized unlabeled data. The varGDLF system is based on mixture model with localized feature selection method. The system has been evaluated on KDD Cup 99 datasets and achieved an accuracy of 85.2%, which means that our detection approach enjoys better accuracy, with 16 features, of 87.33 % and 90.36 for 1NN and SVM respectively.

VII. Conclusion

In this paper, we have proposed an unsupervised feature selection algorithm, which is an enhancement over Laplacian score method. We name our algorithm an Extended Laplacian score, EL in short. More specifically, two main phases are involved in EL during the selection processes. In the first phase, a k-nearest neighbor graph is used to capture the locality preserving power of each feature. In the second phase, a Redundancy Penalization (RP) function is used to eliminate redundancies among selected features. RP is based on the principle of mutual information.

In order to investigate the effectiveness of the proposed method, two intrusion detection systems based on EL and the Laplacian score algorithms are developed. Three different IDS datasets involved in the evaluation processes, the KDD Cup 99, NSL-KDD and Kyoto 2006+ datasets. The performance of EL is compared against the results obtained using Laplacian Score method. Experimental results have shown that our IDS with EL has achieved encouraging results on all datasets and outperformed the Laplacian Score algorithm in terms of classification accuracies.

Although the proposed feature selection algorithm EL has shown good efficiency, it could be further enhanced. For example, adoptive learning algorithms can be used to select an appropriate value for the parameter k. This will be very useful since the proposed method is sensitive to the selection of this parameter. We will put this into consideration to enhance our method.

References

[1] P. Louvieris, N. Clewley, and X. Liu, "Effects-based feature identification for network intrusion detection," *Neurocomputing*, vol. 121, pp. 265–273, 2013.

[2] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Machine Learning: Proceedings of the Eleventh International Conference*, 1994, pp. 121–129.

[3] P. Zhu, W. Zuo, L. Zhang, Q. Hu, and S. C. Shiu, "Unsupervised feature selection by regularized self-representation," *Pattern Recognition*, vol. 48, no. 2, pp. 438–446, 2015.

[4] P. Mitra, C. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," *IEEE transactions on pattern analysis and machine intelligence*, vol. 24, no. 3, pp. 301–312, 2002.

[5] D. Zhang, S. Chen, and Z.-H. Zhou, "Constraint score: A new filter method for feature selection with pairwise constraints," *Pattern Recognition*, vol. 41, no. 5, pp. 1440–1451, 2008.

[6] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Advances in neural information processing systems*, 2005, pp. 507–514.

[7] D. Zhang, J. He, Y. Zhao, Z. Luo, and M. Du, "Global plus local: A complete framework for feature extraction and recognition," *Pattern Recognition*, vol. 47, no. 3, pp. 1433–1442, 2014.

[8] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537–550, 1994.

[9] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.

[10] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[11] N. Kwak and C.-H. Choi, "Input feature selection for classification problems," *IEEE Transactions on Neural Networks*, vol. 13, no. 1, pp. 143–159, 2002.

[12] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[13] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Transactions on Neural Networks*, vol. 20, no. 2, pp. 189–201, 2009.

[14] S. Cang and H. Yu, "Mutual information based input feature selection for classification problems," *Decision Support Systems*, vol. 54, no. 1, pp. 691–698, 2012.

[15] F. Amiri, M. Rezaei Yousefi, C. Lucas, A. Shakery, and N. Yazdani, "Mutual information-based feature selection for intrusion detection systems," *Journal of Network and Computer Applications*, vol. 34, no. 4, pp. 1184–1199, 2011.

[16] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering." in *NIPS*, vol. 14, 2001, pp. 585–591.

[17] X. Niyogi, "Locality preserving projections," in *Neural information processing systems*, vol. 16. MIT, 2004, p. 153.

[18] F. R. Chung, *Spectral graph theory*. American Mathematical Soc., 1997, vol. 92.

[19] A. M. Ambusaidi, X. He, Z. Tan, P. Nanda, L. F. Lu, and T. U. Nagar, "A novel feature selection approach for intrusion detection data classification," in *International Conference on Trust, Security and Privacy in Computing and Communications*. IEEE, 2014, pp. 82–89.

[20] S. J. Stolfo, W. Fan, W. Lee, A. Prodromidis, and P. K. Chan, "Cost-based modeling for fraud and intrusion detection: Results from the jam project," in *DARPA Information Survivability Conference and Exposition*, vol. 2. IEEE, 2000, pp. 130–144.

[21] M. Tavallaee, E. Bagheri, W. Lu, and A.-A. Ghorbani, "A detailed analysis of the kdd cup 99 data set," in *Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications*, 2009, pp. 1–6.

[22] J. Song, H. Takakura, Y. Okabe, M. Eto, D. Inoue, and K. Nakao, "Statistical analysis of honeypot data and building of kyoto 2006+ dataset for nids evaluation," in *Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*. ACM, 2011, pp. 29–36.

[23] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[24] C.-F. Tsai, Y.-F. Hsu, C.-Y. Lin, and W.-Y. Lin, "Intrusion detection by machine learning: A review," *Expert Systems with Applications*, vol. 36, no. 10, pp. 11 994–12 000, 2009.

[25] W. Fan, N. Bouguila, and D. Ziou, "Unsupervised anomaly intrusion detection via localized bayesian feature selection," in *International Conference on Data Mining (ICDM)*. IEEE, 2011, pp. 1032–1037.