Intrusion Detection Using Unsupervised Approach

Jai Puneet Singh(✉) and Nizar Bouguila

CIISE Department, Concordia University, Montreal, Canada
jaipuneet.singh@mail.concordia.ca,
nizar.bouguila@concordia.ca

Abstract. The process of detecting intrusion on network traffic has always remained a key concern for security researchers. During the previous years, intrusion detection had attracted many researchers to find anomaly on NSL-KDD data set. Hence, most of the approaches applied on NSL KDD data set were supervised approaches. We had conducted statistical analysis on this data set using Dirichlet Mixture model. We have seen initialization using Aitchison distance fits better for proportional data. The feature selection highly affects both the performance and results into an improved evaluation of anomaly detection by an unsupervised approach.

Keywords: Mixture models · Intrusion · Aitchison distance · Feature selection

1 Introduction

With an emerging growth of networks and rate of transfer of data through networks has increased the demand for network security. There is a significant literature on Anomaly detection. Anomaly detection deviates from normal traffic and it is important to find an anomaly in an era of communication. Although, there are a lot of articles on intrusion detection, feature selection, and unsupervised learning approach is often underrepresented. There are very limited publicly available data sets for network-based anomaly detection. Earlier KDDCup99 was used heavily for all kind of intrusion detection through machine learning methodology. KDDCup99 has a huge number of redundant records [24]. It was found that around 78% of records in KDDCup99 were duplicated. Mchugh [21] gave many critics on KDDCup dataset and DARPA data set of 1998 as it was not good for applying statistical approaches to learning. The new NSL KDD data set was proposed [2] to overcome the problems present in KDDCup99 and DARPA data sets [1]. NSL KDD data set does not have redundant and duplicates records. There is the lot of work which has been done on NSL KDD data set to find an intrusion [3]. All existing approaches are supervised learning approach. The author in [18] had used Principle component analysis for feature extraction followed by SVM for finding intrusion in NSL KDD data set. The author in [22] had used a combination of classifiers or clusters which are followed by supervised or unsupervised data filtering. The author in [26] had used feature selection technique for a specific group and then comparing corrected KDD data set of feature selection with NSL KDD data set.

F. Belqasmi et al. (Eds.): AFRICATEK 2017, LNICST 206, pp. 192–201, 2018.
https://doi.org/10.1007/978-3-319-67837-5_18

In our paper, we have used unsupervised approach using Dirichlet Mixture Model. The initialization of mixture model is done with K-means using different distance metrics. Aitchison distance metrics shows better results than Euclidean distance for proportional data. It is followed by feature selection on NSL KDD data which reduces features from 41 to 16 features. The comparative analysis has been drawn which, shows that how feature selection and proper initialization increases the detection rate in NSL-KDD data set.

In Sect. 2 of our paper, we have discussed the feature selection approaches and results are showed in form of graph. In Sect. 3, Dirichlet Mixture model is discussed with Aitchison distance being applied on K-means as a distance metrics. Section 4, gives the result of an experiment performed and comparison table. Finally in Sect. 5 concluding remarks are drawn.

2 Feature Selection

There is a subtle difference between feature selection and feature extraction where feature selection performs removal of features which are not relevant when computed with labels during its posterior processes. There are various feature selection methods, popular are being: Stepwise Regression, Stability Selection, Significance Analysis for Microarrays, Weight by Maximum relevance, Least Absolute Selection and Shrinkage Operator (LASSO) etc. Feature extraction transforms the attributes and transformed attributes are a combination of the original attributes. In this process, linear dependence between the features are minimized and projection of original data is on new space. The common feature extraction methods are PCA (principal component analysis), ICA (independent component analysis), Multifactor dimensionality reduction, Latent semantic analysis etc. The novel methods of feature extraction on proportional data were proposed by an author in [20] which extracts features of proportional data using data separation by Dirichlet distribution. In our paper, we have concentrated upon feature selection which is different from feature extraction.

2.1 Weight by Maximum Relevance

It has been proposed by Blum et al. [4], is a filter that measures the dependence between every feature x and the classification feature y (i.e., the label) using Pearson's linear correlation, F-test scores, and mutual information [4, 19]. The high score by mutual correlation reveals the features which are important. The NSL KDD Dataset has 41 features and in order to reduce the complexity and finding an optimal solution we have reduced to 16 features taking into an account that Weight by Maximum Relevance score of the feature is $f \geq 0.05$. The output obtained can be shown by the Fig. 1.

Weight by Maximum Relevance correlation vector can be defined by Pearson Correlation coefficient as:

$$R(i) = \frac{cov(X_i, Y)}{\sqrt{Var(Xi)Var(Y)}} \tag{1}$$

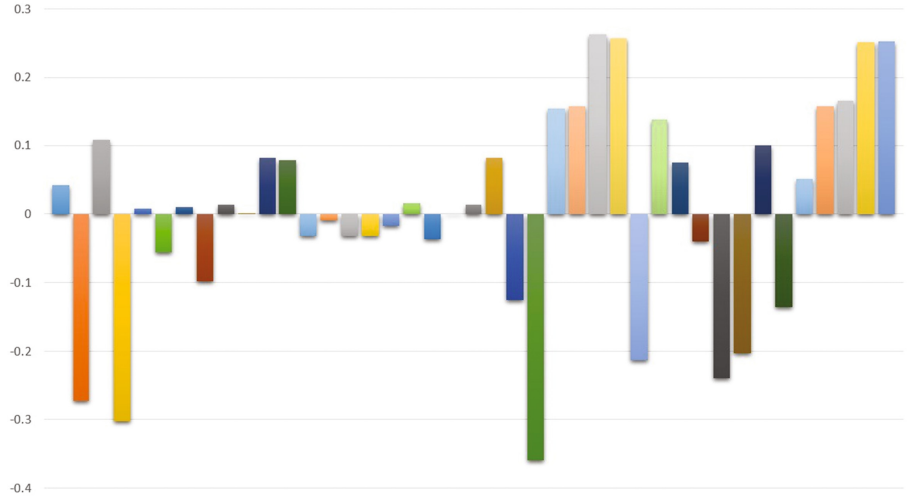

Fig. 1. Score obtained after applying weight by maximum relevance feature selection technique

The equation can be written as:

$$R(i) = \frac{\sum_{k=1}^{M} (x_{k,i} - \bar{x})(yk - \bar{y})}{\sqrt{\sum_{k=1}^{M} (x_{k,i} - \bar{x})^2 \sum_{k=1}^{M} (yk - \bar{y})^2}} \tag{2}$$

This can only detect the linear dependency between variable and target [17].

2.2 Least Absolute Selection and Shrinkage Operator (LASSO)

Tibshirani Robert [25] explains feature selection by checking vector β which is a coefficient vector. It minimizes the residual sum of squares which is related to coefficient being less. It shrinks coefficients and set others to zero, therefore tries to retain the good features of both subset selection and ridge regression. It is given $(x_1, x_2, \ldots, x_D)$ and an outcome be y, the LASSO should fit linear model. The computation of LASSO is a quadratic problem and can be solved by standard numerical analysis algorithms. LASSO does shrinkage and variable selection whereas ridge regression only shrinks. The initial idea is to start working with the large value of λ and slowly start decreasing it. The minimization for LASSO can be expressed as follow:

$$\sum_{i=1}^{n} (y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3}$$

In this equation yi is the outcome variable, for cases $i = 1, 2, \ldots, n$ features $x_{ij}, j = 1, 2, \ldots, p$. Figure 2 represents feature selection by LASSO and reducing features to 16 features by taking into an account $f \geq 0.0053$.

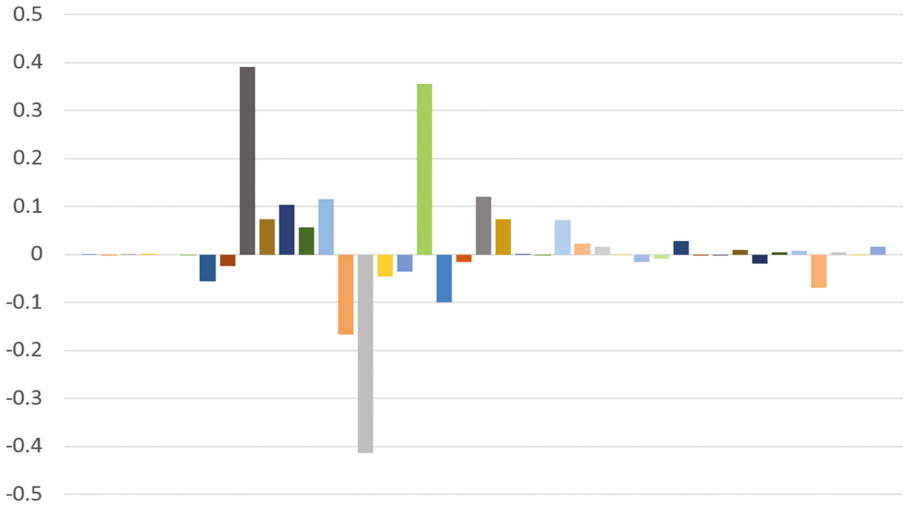

Fig. 2. Score obtained after applying LASSO feature selection technique

3 Proposed Method

Let $X = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N\}$ be the data set with N D-dimensional such that Dirichlet mixture model being applied on it. The density function of Dirichlet mixture model can be given by

$$p(X_i|\theta) = \sum_{j=1}^{M} p_j p(X_i|\alpha_j) \tag{4}$$

where α_j is the parameter vector of component j, p_j is the mixing proportion which should be positive and always sum to 1. $\theta = \{p_1, p_2, \ldots, p_M; \boldsymbol{\alpha_1}, \boldsymbol{\alpha_2}, \ldots, \boldsymbol{\alpha_M}\}$ is the complete set of parameters fully characterizing the mixture $M \geq 1$ is the number of components. Each Dirichlet distribution can be written in the form

$$p(X_i, a_j) = \frac{1}{\beta(\alpha)} \prod_{d=1}^{D} X_{id}^{\alpha jd-1} \tag{5}$$

$$\beta(\alpha) \frac{\prod_{d=1}^{D} \Gamma(\alpha_{jd})}{\Gamma(\sum_{d=1}^{D} \alpha_{jd})} \tag{6}$$

where $x_{id} > 0$ $d = 1, 2, \ldots, D, X = \{X_{i1} + X_{i2}, \ldots, +X_{id} = 1\}$ and $\alpha_j = (\alpha_{j1}, \alpha_{j2} \ldots \alpha_{jD})$ represents parameter vector for j^{th} population. Let N D-dimensional vector be $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$ a data set of with a common, but unknown, probability density function $p(\mathbf{X}_i|\theta)$ as given in above equation.

We supposed that the number of mixtures component is known. The ML estimation method consist of getting the mixture parameters that maximize log likelihood function. The below equation defines the posterior probability obtained after solving log likelihood function. This function is used in as an E-step of Expectation Maximization (EM) algorithm.

$$p(j|\mathrm{X_i},\ \alpha_\mathrm{j}) = \frac{p_j p(\mathrm{X_i}|\alpha_\mathrm{j})}{\sum_{k=1}^{K} p_{ikp}(\mathrm{X_i}|\alpha_\mathrm{k})} \tag{7}$$

Now, using this expectation our goal is to maximize complete log likelihood. During the process we also have to ensure that constraint $p_j \geq 0$ as well as $\sum_{j=1}^{M} p_j = 1$. In maximization step of the algorithm, we have to update the parameters α until it converges to get the best result. As it is to be noted that closed form solution of α does not exist. In the maximization step, the iterative approach of Newton Raphson method has been used as explained by the author in [11] for estimation of α parameters.

During the initialization of parameters for Dirichlet mixture model, we use K-means algorithm as given in Algorithm 1 to initialize the parameters. We have compared our results by changing K-means algorithm using different distance metrics. We have used Euclidean distance and Aitchison distance inside K-means for initialization of parameters for Dirichlet mixture model. As we know that Aitchison distance outperforms euclidean distance metrics when proportional data is in question. In order to increase the performance of an algorithm, we have used feature selection methodology [5, 7, 9, 15, 16]. In order to perform feature selection, the first step we have taken to normalize the NSL KDD data set using Eq. 8.

$$x_i = \frac{x_i}{x_1 + x_2 \ldots + x_D} \tag{8}$$

After obtaining proportional data, which act as an input for Weight by Maximum Relevance (*WMR*) proposed by Blum et al. [4] and Least Absolute Selection and Shrinkage Operator (*LASSO*) for selection of features from a data set.

Normalization of data leads vector to $(X_{i1} + X_{i2}, \ldots X_{iD} = 1)$ unit sum constraint and each $X_i \geq 0$. After normalization, we have used Dirichlet Mixture Model with an initialization of parameters using K-means with Aitchison and Euclidean distance metrics.

In Algorithm 1, the distance metric which has been used is Aitchison Distance metric which can be given as:

$$d_{AD}(x, y) = \frac{1}{D} \sum_{i<j} \left(\log \frac{x_i}{x_j} - \log \frac{y_i}{y_j} \right) \tag{9}$$

Algorithm 1. K-Means Algorithm

1: Set the Initial number of centroids randomly or sequentially
2: Calculate the distance between each data point and cluster centers
3: **repeat**:
4: Assign the minimum **distance data points** to cluster center whose distance is minimum to that point.
5: Recalculate the cluster center using:
6: $c_i = \frac{1}{m_i} \sum_{j=1}^{m_i} x\,(i)$; m_i represents total number of data points in a cluster
7: Re-calculate the distance between each data point and newly obtained cluster center
8: **until** : No data point is reassigned.

Algorithm 2. EM Algorithm Dirichlet Mixture Model

1: **Input:** Data set $(\mathbf{X}_1 + \mathbf{X}_2 ... \mathbf{X}_N)$ and specified number of components M.
2: Apply the k-means algorithm as given in Algorithm 1 on N D-dimensional vectors to obtain initial M clusters.
3: calculate $p_j = \dfrac{\text{Number of elements in class j}}{\text{N}}$
4: Apply moments method to obtain α parameters.
5: **Expectation-Maximization step** after Initialization
6: E-Step: Compute the posterior probability $p\,(j|\mathbf{X_i}, \boldsymbol{\alpha})$
7: M-Step:
8: **repeat**:
9: Update priors p_j using equation 7 .
10: Update the parameters α using Newton Raphson method.[11].
11: **until** : $p_j \leq \epsilon$, discard j and go to E-Step.
12: if convergence test is passed then terminate, else go to E-Step.

$$d^2_{AD}(x, y) = \sum_{k=1}^{D} \left(\log \frac{x_i}{g(x_j)} - \log \frac{y_i}{g(y_j)} \right) \tag{10}$$

The methodology used in our experiment is as follows:

4 Experiment with NSL KDD Data Set

We have taken NSL KDD 2009 data-set for performing Intrusion detection. The NSL KDD data set contains 41 features and data set contains normal and attack sets. The attacks can be divided into four parts which are: Denial of Service Attack (DoS), User to Root attack (U2R), Remote to local attack (R2L) and probing attack and rest are normal sets. In our experiment, we have taken only normal and attack sets into consideration without finding different types of attacks. In our methodology, we have used Dirichlet Mixture Model for clustering of a data set which contains 41 features. While

performing clustering using Dirichlet mixture model results into 51.12% of accuracy which was relatively increased to 53.44% when clustering was performed with initialization of k-means using Aitchison distance $d_{AD}(x,y) = \frac{1}{D}\sum_{i<j}\left(\log\frac{x_i}{x_j} - \log\frac{y_i}{y_j}\right)$ instead of Euclidean distance in k-means algorithm. In our experiments, we have done feature selection using the methodology of Weight by maximum relevance where features were reduced to 16 features. The experiment on 16 features using Dirichlet Mixture model with euclidean distance in K-means during initialization results into 52.54% of accuracy and 56.37% was obtained when initialization was done with K-means using Aitchison distance in Dirichlet mixture model as seen in Table 1 and Fig. 3. To depict our results, we have used confusion matrix as shown in Figs. 4 and 5. to show the accuracy of our results. Accuracy is defined as the percentage of correctly classified vectors. The accuracy of results can be written as:

Algorithm 3

1: **Input:** The Data (X_i) with labels.
2: To Normalize the data using equation 8
3: To find the correlation between data and labels using Weight by Maximum Relevance **or**
4: To find least square regression coefficients using set of regularization coefficients Lambda.
5: To select attributes from data set by using figure 1 or figure 2.
6: **Output:** Dimensionally reduced data set (X_i)
7: **Next Process**
8: **Input** To input obtained data without using Labels to Algorithm 2 with number of mixtures.
9: **Output**: We get clusters of normal data and anomaly data.

$$Accuracy = 100 \times \frac{\text{Correctly identified vector}}{\text{total vectors}} \tag{11}$$

Table 1. Accuracy, Precision and Sensitivity obtained after applying different methods

S.No.	Process	Accuracy	Precision	Sensitivity
1.	DMM (Euclidean Distance)	51.12%	0.78	0.55
2.	DMM (Aitchison Distance)	53.44%	0.76	0.56
3.	FS WMR DMM (Euclidean Distance)	52.54%	0.78	0.58
4.	FS WMR DMM (Aitchison Distance)	56.37%	0.80	0.57

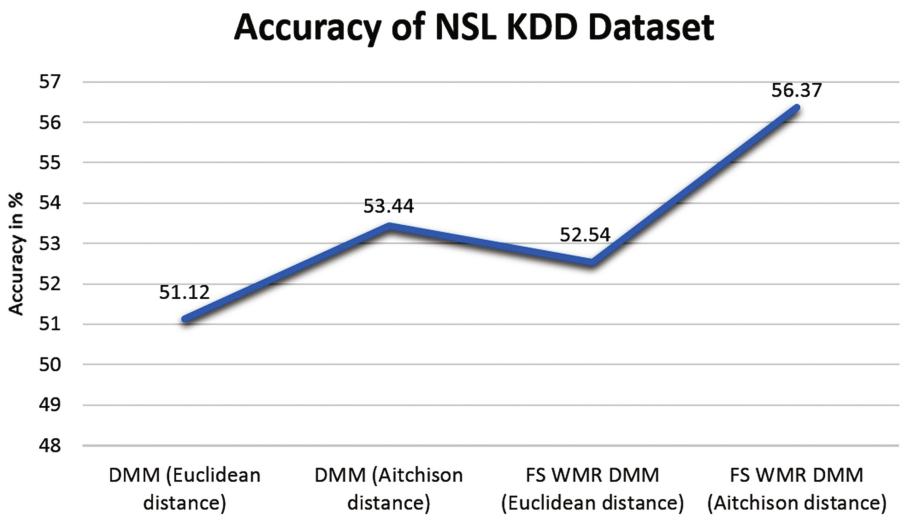

Fig. 3. Accuracy of DMM models using different techniques

	Yes	No
Yes	5386	1492
No	4300	672

DMM (Euclidean distance)

	Yes	No
Yes	4953	1564
No	3953	1380

DMM (Aitchison distance)

Fig. 4. Confusion matrix of DMM, initialization with K-means euclidean and Aitchison distance

	Yes	No
Yes	5587	1513
No	4111	639

FS WMR DMM (Euclidean distance)

	Yes	No
Yes	5128	1285
No	3884	1553

FS WMR DMM (Aitchison distance)

Fig. 5. Confusion matrix of DMM after feature selection, initialization with K-means Euclidean and Aitchison distance

In our case, we have used only test data without labels. Our results are better than SVM approach where accuracy determined is 51.90% [19] where the model was trained before determining the intrusions. The author in [14] obtained results in one of the clustering method was 47% which is comparably less than our approach.

5 Conclusion

In this paper, we have statistically analyzed the entire NSL KDD data set. The analysis showed that initialization by K-means using Aitchison distance on proportional data improves the accuracy of the model. It shows that improving the initialization of a mixture model gives the better result. Every data set is normalized before performing an unsupervised algorithm which leads to proportional data. The proportional data is well handled with Aitchison distance. The limitation of above method is that it is computationally expensive process and further research can be taken place for optimization of this current technique. In NSL KDD data set, 16 features had shown strong contribution for anomaly detection. Finally, we got better results than previous unsupervised approaches. Our basis is to state the baseline for unsupervised learning for future IDS solution.

Acknowledgment. The authors would like to thank Mitacs Inc. for providing the research support and Concordia University for providing the facilities to conduct a research.

References

1. Darpa intrusion detection evaluation. http://www.ll.mit.edu/IST/ideval/data/dataindex.html. Accessed 05 Nov 2016
2. NSL-KDD data set for network-based intrusion detection systems. http://nsl.cs.unb.ca/KDD/NSLKDD.html. Accessed 05 Nov 2016
3. Al-Yaseen, W.L., Othman, Z.A., Nazri, M.Z.A.: Multi-level hybrid support vector machine and extreme learning machine based on modified k-means for intrusion detection system. Expert Syst. Appl. **67**, 296–303 (2017)
4. Blum, A.L., Langley, P.: Selection of relevant features and examples in machine learning. Artif. Intell. **97**(1), 245–271 (1997)
5. Bouguila, N.: Bayesian hybrid generative discriminative learning based on finite Liouville mixture models. Pattern Recogn. **44**(6), 1183–1200 (2011)
6. Bouguila, N., ElGuebaly, W.: Discrete data clustering using finite mixture models. Pattern Recogn. **42**(1), 33–42 (2009)
7. Bouguila, N., Ziou, D.: MML-based approach for finite Dirichlet mixture estimation and selection. In: Perner, P., Imiya, A. (eds.) MLDM 2005. LNCS, vol. 3587, pp. 42–51. Springer, Heidelberg (2005). doi:10.1007/11510888_5
8. Bouguila, N., Ziou, D.: On fitting finite Dirichlet mixture using ECM and MML. In: Singh, S., Singh, M., Apte, C., Perner, P. (eds.) ICAPR 2005. LNCS, vol. 3686, pp. 172–182. Springer, Heidelberg (2005). doi:10.1007/11551188_19
9. Bouguila, N., Ziou, D.: A countably infinite mixture model for clustering and feature selection. Knowl. Inf. Syst. **33**(2), 351–370 (2012)
10. Bouguila, N., Ziou, D., Hammoud, R.I.: On Bayesian analysis of a finite generalized Dirichlet mixture via a metropolis-within-gibbs sampling. Pattern Anal. Appl. **12**(2), 151–166 (2009)

11. Bouguila, N., Ziou, D., Vaillancourt, J.: Novel mixtures based on the Dirichlet distribution: application to data and image classification. In: Perner, P., Rosenfeld, A. (eds.) MLDM 2003. LNCS, vol. 2734, pp. 172–181. Springer, Heidelberg (2003). doi:10.1007/3-540-45065-3_15
12. Elguebaly, T., Bouguila, N.: Finite asymmetric generalized Gaussian mixture models learning for infrared object detection. Comput. Vis. Image Underst. **117**(12), 1659–1671 (2013)
13. Epaillard, E., Bouguila, N.: Proportional data modeling with hidden Markov models based on generalized dirichlet and beta-liouville mixtures applied to anomaly detection in public areas. Pattern Recogn. **55**, 125–136 (2016)
14. Eskin, E., Arnold, A., Prerau, M., Portnoy, L., Stolfo, S.: A geometric framework for unsupervised anomaly detection. In: Barbará, D., Jajodia, S. (eds.) Applications of Data Mining in Computer Security. Advances in Information Security, vol. 6, pp. 77–101. Springer, Boston (2002)
15. Fan, W., Bouguila, N., Ziou, D.: Unsupervised anomaly intrusion detection via localized Bayesian feature selection. In: 2011 IEEE 11th International Conference on Data Mining (ICDM), pp. 1032–1037. IEEE (2011)
16. Fan, W., Bouguila, N., Ziou, D.: Unsupervised hybrid feature extraction selection for high-dimensional non-gaussian data clustering with variational inference. IEEE Trans. Knowl. Data Eng. **25**(7), 1670–1685 (2013)
17. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
18. Heba, F.E., Darwish, A., Hassanien, A.E., Abraham, A.: Principle components analysis and support vector machine based intrusion detection system. In: 2010 Proceedings of the 10th International Conference on Intelligent Systems Design and Applications, pp. 363–367. IEEE (2010)
19. Iglesias, F., Zseby, T.: Analysis of network traffic features for anomaly detection. Mach. Learn. **101**(1–3), 59–84 (2015)
20. Masoudimansour, W., Bouguila, N.: Dimensionality reduction of proportional data through data separation using Dirichlet distribution. In: Kamel, M., Campilho, A. (eds.) ICIAR 2015. LNCS, vol. 9164, pp. 141–149. Springer, Cham (2015). doi:10.1007/978-3-319-20801-5_15
21. McHugh, J.: Testing intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln laboratory. ACM Trans. Inf. Syst. Secur. (TISSEC) **3**(4), 262–294 (2000)
22. Panda, M., Abraham, A., Patra, M.R.: A hybrid intelligent approach for network intrusion detection. Procedia Eng. **30**, 1–9 (2012)
23. Singh, S., Singh, M., Apte, C., Perner, P.: Pattern Recognition and Data Mining: Third International Conference on Advances in Pattern Recognition, ICAR 2005, Bath, UK, 22–25 August 2005, vol. 3686. Springer, Heidelberg (2005)
24. Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A.A.: A detailed analysis of the KDD CUP 99 data set. In: 2009 Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications (2009)
25. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc. Ser. B (Methodol.) **58**, 267–288 (1996)
26. Zargari, S., Voorhis, D.: Feature selection in the corrected KDD-dataset. In: 2012 Third International Conference on Emerging Intelligent Data and Web Technologies (EIDWT), pp. 174–180. IEEE (2012)