A Novel Anomaly Detection System Based on HFR-MLR Method

Eunhye Kim[1] and Sehun Kim[2]

[1] IT Convergence Technology Research Division, ETRI, South Korea
eunhye@etri.re.kr
[2] Internet Security Lab., KAIST, South Korea
shkim@kaist.ac.kr

Abstract. Reducing the data space and then classifying anomalies based on the reduced feature space is vital to real-time intrusion detection. In this study, a novel framework is developed for logistic regression-based anomaly detection and hierarchical feature reduction (HFR) to preprocess network traffic data before detection model training. The proposed dimensionality reduction algorithm optimally excludes the redundancy of features by considering the similarity of feature responses through a clustering analysis based on the feature space reduced by factor analysis, thus helping to rank the importance of input features (essential, secondary and insignificant) with low time complexity. Classification of anomalies over the reduced feature space is based on a multinomial logistic regression (MLR) model to detect multi-category attacks as an outcome with the goal of reinforcing detection efficiency. The proposed system not only achieves a significant detection performance, but also enables fast detection of multi-category attacks.

Keywords: Anomaly detection, Dimensionality reduction, Hierarchical clustering, Multinomial logistic regression.

1 Introduction

As the potential damage caused by malicious network activities has become more serious, the need to defend against these threats has increased significantly. The network intrusion detection system (NIDS), as a vital system in the network security infrastructure, aims to detect attacks quickly and accurately; its role is becoming more important. To achieve this objective, previously observed attack patterns need to be analyzed and profiled so that criteria for what constitutes normal traffic or an attack can be determined and applied to newly captured patterns for intrusion detection. In the detection approaches of NIDS, many studies have applied data mining techniques such as a support vector machine (SVM) and neural networks [1-2].

Although the techniques applied in previous works have shown good results in terms of data classification, they are not favorable for large-scale datasets because the training complexity is very much dependent on the amount of data in the training set. Especially, some data features in the classifiers used in NIDS may be redundant or

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*,
Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_43, © Springer-Verlag Berlin Heidelberg 2014

may contribute little to the detection process. Extraneous features and the complex relationships that exist among the features can make it harder to detect suspicious behavior patterns and can increase the computation time. Therefore, through feature dimensionality reduction, NIDS must reduce the amount of data to be processed for computationally efficient and effective detection.

This study proposes a multinomial logistic regression (MLR)-based network anomaly detection system based on hierarchical feature reduction (HFR) to preprocess network traffic data before detection model training. The proposed HFR algorithm optimally excludes the redundancy of features by considering the similarity of feature responses through a clustering analysis based on the feature space reduced by factor analysis. The performance of the proposed method is evaluated using different data sets reduced by the ranking of the importance of input features. Classification of intrusions over the reduced feature space was based on the MLR model, a method well suited for analyzing multi-type outcomes with high speed in learning techniques. Our classification model was developed for the detection of multi-category attacks as an outcome to reinforce detection efficiency, unlike previous studies that were focused on a binary outcome (e.g., normal or abnormal). The experiment with the NSL-KDD dataset showed a significant detection rate through a good subset of features with a significant improvement in speed.

This paper is organized as follows. In Section 2, several examples of related work are reviewed. The proposed algorithm is then described in Section 3. Section 4 gives details of the experiments as well as the results. The study is concluded with a summary and plans for future research in Section 5.

2 Related Work

Anomaly detection depends on the idea that the characteristics of normal behavior can be distinguished from those of abnormal behavior. Statistical modeling remains the most common approach to anomaly intrusion detection; this method includes cluster analysis, Bayesian analysis, principal component analysis, and the fuzzy inference approach. Leung et al. [3] carried out research based on density and a grid-based clustering method for anomaly detection. Chan et al. [4] investigated both the distance and density of clusters and found that attacks were often in outlying clusters with statistically low or high densities. Valdes et al. [5] employed naive Bayesian networks to perform intrusion detection on traffic bursts. Xu et al. [6] used continuous time Bayesian networks and avoided specifying a fixed update interval common to discrete-time models. Huang et al. [7] presented a simple algorithmic framework for network-wide anomaly detection that relies on distributed tracking combined with approximate PCA. Toosi et al. [8] combined a neuro-fuzzy network, the fuzzy inference approach, and genetic algorithms to design an intrusion detection system.

Most previous studies were conducted based on all possible independent variables. Unnecessary variables can create bias and lead the model either to overestimate or underestimate the detecting values. In this study, in order to reduce the amount of training data, the HFR method was developed using unsupervised data mining techniques, and applied before MLR training. Also, our classification model was developed for the detection of multi-category attacks as an outcome to reinforce detection efficiency, unlike previous studies that were focused on a binary outcome.

3 Proposed Framework

The proposed framework consists of three main phases. In the first phase, the feature redundancy can be reduced by considering the similarity of variable-responses to the training data set through clustering analysis. The proposed scheme can hierarchically reduce the features, thus helping to rank the importance of input features. Then, in the second phase an anomaly detection model using MLR is constructed with the reduced training dataset resulted from the feature reduction algorithm. As a result of the model, the odds ratios provide an estimate of the likelihood of being identified as an anomaly. In the third phase, test data are used to detect anomalies according to attack types based on the developed MLR model. The performance of our anomaly detection model is evaluated using a cross validation testing concept.

3.1 Proposed Hierarchical Feature Reduction

Feature reduction involves processes of determining the evidence that can be taken from the raw data that is most useful for analysis. To exclude the redundancy of features and to improve the performance of classification, statistical techniques are used, including factor analysis which is one of the most widely used dimensionality reduction techniques and hierarchical clustering which does not require predetermined numbers of groups and has the advantage of low time complexity.

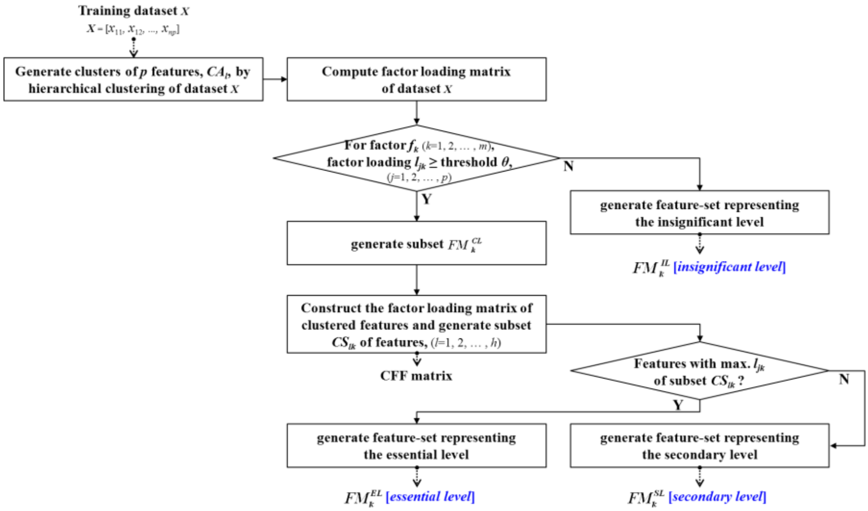
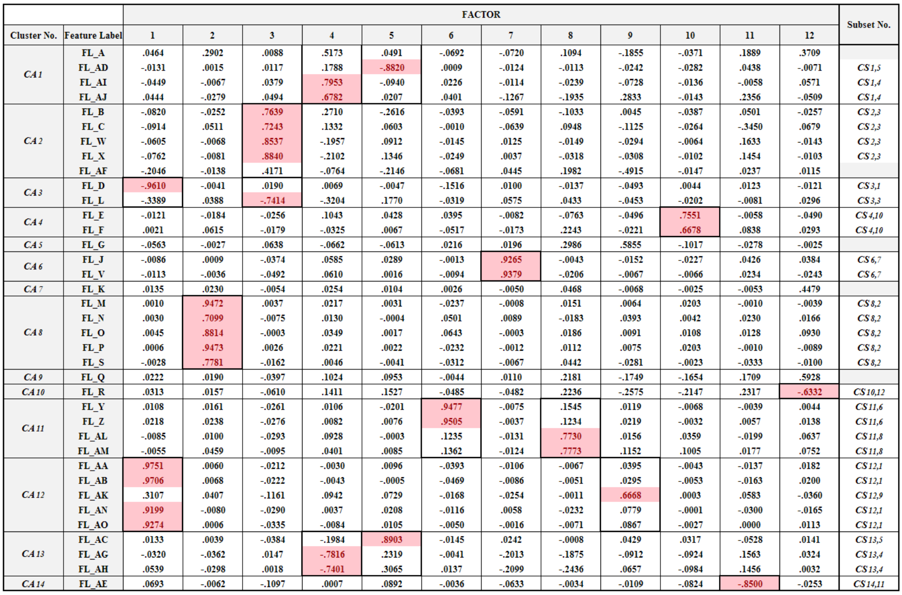

Training dataset X
$X = [x_{11}, x_{12}, ..., x_{np}]$
Generate clusters of p features, CA_l, by hierarchical clustering of dataset X
Compute factor loading matrix of dataset X
For factor f_k (k=1, 2, … , m), factor loading $l_{jk} \geq$ threshold θ, (j=1, 2, … , p)
N
Y
generate feature-set representing the insignificant level
FM_k^{IL} [insignificant level]
generate subset FM_k^{CL}
Construct the factor loading matrix of clustered features and generate subset CS_{lk} of features, (l=1, 2, … , h)
CFF matrix
Features with max. l_{jk} of subset CS_{lk} ?
N
Y
generate feature-set representing the essential level
generate feature-set representing the secondary level
FM_k^{EL} [essential level]
FM_k^{SL} [secondary level]

Fig. 1. Proposed HFR procedure

In an application of factor analysis, if feature dimensionality reduction is based on only the degree of contribution induced from observing which variables are most heavily loaded on certain factors, the selected features may be redundant as the information that they include is contained in other features. This redundancy can be reduced by considering the similarity of variable-responses to the training data set through clustering analysis.

Therefore, a hierarchical feature dimensionality reduction algorithm is proposed in which the factor analysis and hierarchical clustering are combined. In the proposed algorithm, hierarchical clustering is initially applied and factor analysis is then applied to the training data set, as shown in Fig. 1. Based on *h* clusters of features through hierarchical clustering, features are extracted in which the factor loadings are higher than a certain threshold (subset $\boldsymbol{CS_{lk}}$ in Fig. 1). The redundancy of features with a high value of factor loadings is then reduced if they are in the same cluster, which organizes a good subset (subset $\boldsymbol{FM_k^{EL}}$) of features critical to the performance of classifiers. The strongest point of the proposed feature reduction scheme is that this method can hierarchically reduce the features, thus helping to rank the importance of the input features (essential, secondary and insignificant) with low time complexity. Using far fewer instances, the proposed method can produce high quality datasets that sufficiently represent all of the instances in the original dataset. The clustered feature-factor (CFF) matrix generating subset of the significant features is shown in Fig. 2.



		FACTOR												Subset No.
Cluster No.	Feature Label	1	2	3	4	5	6	7	8	9	10	11	12	
CA 1	FL_A	.0464	.2902	.0088	.5173	.0491	-.0692	-.0720	.1094	-.1855	-.0371	.1889	.3709	
	FL_AD	-.0131	.0015	.0117	.1788	-.8820	.0009	-.0124	-.0113	-.0242	-.0282	.0438	-.0071	$CS_{1,5}$
	FL_AI	-.0449	-.0067	.0379	.7953	-.0940	.0226	-.0114	-.0239	-.0728	-.0136	-.0058	.0571	$CS_{1,4}$
	FL_AJ	.0444	-.0279	.0494	.6782	.0207	.0401	-.1267	-.1935	.2833	-.0143	.2356	-.0509	$CS_{1,4}$
CA 2	FL_B	-.0820	-.0252	.7639	.2710	-.2616	-.0393	-.0591	-.1033	.0045	-.0387	.0501	-.0257	$CS_{2,3}$
	FL_C	-.0914	.0511	.7243	.1332	.0603	-.0010	-.0639	.0948	-.1125	-.0264	-.3450	.0679	$CS_{2,3}$
	FL_W	-.0605	-.0068	.8537	-.1957	.0912	-.0145	.0125	-.0149	-.0294	-.0064	.1633	-.0143	$CS_{2,3}$
	FL_X	-.0762	-.0081	.8840	-.2102	.1346	-.0249	.0037	-.0318	-.0308	-.0102	.1454	-.0103	$CS_{2,3}$
	FL_AF	-.2046	-.0138	.4171	-.0764	-.2146	-.0681	.0445	.1982	-.4915	-.0147	.0237	.0115	
CA 3	FL_D	-.9610	-.0041	.0190	.0069	-.0047	-.1516	.0100	-.0137	-.0493	.0044	.0123	-.0121	$CS_{3,1}$
	FL_L	-.3389	.0388	-.7414	-.3204	.1770	-.0319	.0575	.0433	-.0453	-.0202	-.0081	.0296	$CS_{3,3}$
CA 4	FL_E	-.0121	-.0184	-.0256	.1043	.0428	.0395	-.0082	-.0763	-.0496	.7551	-.0058	-.0490	$CS_{4,10}$
	FL_F	.0021	.0615	-.0179	-.0325	.0067	-.0517	-.0173	.2243	-.0221	.6678	.0838	.0293	$CS_{4,10}$
CA 5	FL_G	-.0563	-.0027	.0638	-.0662	-.0613	.0216	.0196	.2986	.5855	-.1017	-.0278	-.0025	
CA 6	FL_J	-.0086	.0009	-.0374	.0585	.0289	-.0013	.9265	-.0043	-.0152	-.0227	.0426	.0384	$CS_{6,7}$
	FL_V	-.0113	-.0036	-.0492	.0610	.0016	-.0094	.9379	-.0206	-.0067	-.0066	.0234	-.0243	$CS_{6,7}$
CA 7	FL_K	.0135	.0230	-.0054	.0254	.0104	.0026	-.0050	.0468	-.0068	-.0025	-.0053	.4479	
CA 8	FL_M	.0010	.9472	.0037	.0217	.0031	-.0237	-.0008	.0151	.0064	.0203	-.0010	-.0039	$CS_{8,2}$
	FL_N	.0030	.7099	-.0075	.0130	-.0004	.0501	.0089	-.0183	.0393	.0042	.0230	.0166	$CS_{8,2}$
	FL_O	.0045	.8814	-.0003	.0349	.0017	.0643	-.0003	.0186	.0091	.0108	.0128	.0930	$CS_{8,2}$
	FL_P	.0006	.9473	.0026	.0221	.0022	-.0232	-.0012	.0112	.0075	.0203	-.0010	-.0089	$CS_{8,2}$
	FL_S	-.0028	.7781	-.0162	.0046	-.0041	-.0312	-.0067	.0442	-.0281	-.0023	-.0333	-.0100	$CS_{8,2}$
CA 9	FL_Q	.0222	.0190	-.0397	.1024	.0953	-.0044	.0110	.2181	-.1749	-.1654	.1709	.5928	
CA 10	FL_R	.0313	.0157	-.0610	.1411	.1527	-.0485	-.0482	.2236	-.2575	-.2147	.2317	-.6332	$CS_{10,12}$
CA 11	FL_Y	.0108	.0161	-.0261	.0106	-.0201	.9477	-.0075	.1545	.0119	-.0068	-.0039	.0044	$CS_{11,6}$
	FL_Z	.0218	.0238	-.0276	.0082	.0076	.9505	-.0037	.1234	.0219	-.0032	.0057	.0138	$CS_{11,6}$
	FL_AL	-.0085	.0100	-.0293	.0928	-.0003	.1235	-.0131	.7730	.0156	.0359	-.0199	.0637	$CS_{11,8}$
	FL_AM	-.0055	.0459	-.0095	.0401	.0085	.1362	-.0124	.7773	.1152	.1005	.0177	.0752	$CS_{11,8}$
CA 12	FL_AA	.9751	.0060	-.0212	-.0030	.0096	-.0393	-.0106	-.0067	.0395	-.0043	-.0137	.0182	$CS_{12,1}$
	FL_AB	.9706	.0068	-.0222	-.0043	-.0005	-.0469	-.0086	-.0051	.0295	-.0053	-.0163	.0200	$CS_{12,1}$
	FL_AK	.3107	.0407	-.1161	.0942	.0729	-.0168	-.0254	-.0011	.6668	.0003	.0583	-.0360	$CS_{12,9}$
	FL_AN	.9199	-.0080	-.0290	.0037	.0208	-.0116	.0058	-.0232	.0779	-.0001	-.0300	-.0165	$CS_{12,1}$
	FL_AO	.9274	.0006	-.0335	-.0084	.0105	-.0050	-.0016	-.0071	.0867	-.0027	.0000	.0113	$CS_{12,1}$
CA 13	FL_AC	.0133	.0039	-.0384	-.1984	.8903	-.0145	.0242	-.0008	.0429	.0317	-.0528	.0141	$CS_{13,5}$
	FL_AG	-.0320	-.0362	.0147	-.7816	.2319	-.0041	-.2013	-.1875	-.0912	-.0924	.1563	.0324	$CS_{13,4}$
	FL_AH	.0539	-.0298	.0018	-.7401	.3065	.0137	-.2099	-.2436	.0657	-.0984	.1456	.0032	$CS_{13,4}$
CA 14	FL_AE	.0693	-.0062	-.1097	.0007	.0892	-.0036	-.0633	-.0034	-.0109	-.0824	-.8500	-.0253	$CS_{14,11}$

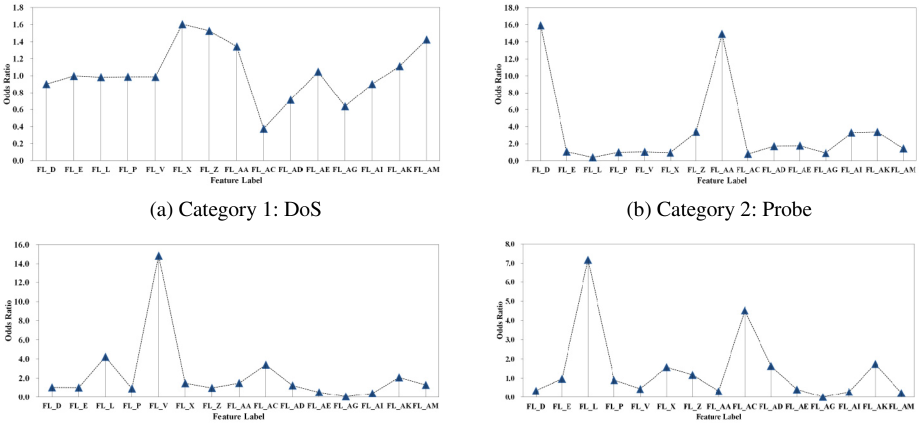
Fig. 2. Clustered feature-factor matrix of the training dataset

3.2 Anomaly Detection Method

The proposed anomaly detection method uses an MLR to build a classifier model. Unlike a binary logistic model, in which a dependent variable has only a binary choice, the dependent variable in the MLR model can have more than two choices that are coded categorically, and one of the categories is taken as the reference category [9]. This study used ‘0’ (normal) as the reference category. Suppose Y_i is the dependent variable with five categories for individual connection i; the probability of being in category m can be represented with the chosen reference category:

$$\log\frac{P(Y_i=m)}{P(Y_i=0)}=\alpha_m+\sum_{k=1}^{p}\beta_{mk}x_{ik}=Z_{mi} \tag{1}$$

where m=‘1’ [DoS], ‘2’ [Probe], ‘3’ [R2L], and ‘4’ [U2R]. Our MLR modeling is performed with a significance threshold of 0.05 for adding variables and an insignificance threshold of 0.1 for removing variables, yielding a set of variables that are associated with the outcome in a statistically significant way. The final MLR model calculates the predicted probabilities of being in the outcome category for each connection record; the classification of the unordered set {0, 1, 2, 3, 4} is conducted on the basis of that probability. The odds ratio of the proposed MLR model, consisting of the essential level variables, for detecting each attack relative to the normal category is shown in Fig. 3.



(a) Category 1: DoS

(b) Category 2: Probe

(c) Category 3: R2L

(d) Category 4: U2R

Fig. 3. Odds ratio of the essential features by attack category

4 Experiments and Results

The data used for testing is NSL-KDD, which is a new dataset for the evaluation of studies in network intrusion detection systems [10]. Each NSL-KDD connection record contains 41 features (e.g., protocol type, service, and flag) and is labeled as either normal or an attack, with one specific attack type. The attacks fall into one of the four categories: DoS, Probe, R2L, and U2R. The NSL-KDD training set contains a total of 22 training attack types, with an additional 17 types in the test set. In the experiments, the dataset was partitioned into subsets. The training set contained 125,973 records and five evaluation sets were comprised of 111,630 records.

We compared the evaluation results of models using a selected feature-set in the essential level with those using essential and secondary level features by the proposed feature reduction algorithm. As can be seen in Fig. 4, the classification rates using the feature-set in the essential level are comparable to those using the essential and secondary levels, except for the case of the Probe class. Both sets of performance results show difficulties in detecting R2L attacks, which are embedded in the data packets themselves and do not form a sequential pattern. These were assigned to incorrect classes and lowered the detection rate. And, with the too small number of instances of U2R attacks in the NSL-KDD dataset, both models of reduced features provided relatively low performance for the U2R class. However, compared with the results of the performance with essential and secondary level features, test numbers 1, 6, 10, 11, 13, and 14 showed higher detection rates with lower false alarm rates, as shown in Fig. 4. It can be said that the proposed HFR method achieves significant detection rates that demonstrate the possibility of successfully detecting attacks with a significant improvement in speed by using only a half percent of the comparison feature-set and 39.0% as compared with the full feature-set. Our method also improved detection times by 23.8% compared to those including the secondary level features.

Experiments were also attempted to evaluate the performance of our anomaly detection scheme compared with that of several other methods; results are shown in Table 1. It can be stated that all the algorithms tested on the KDD data set offered an acceptable level of detection performance for Normal, DoS and Probe classes; they did not have good performance on R2L and U2R attacks. The SVM with BIRCH clustering [11] and ESC-IDS [8] showed the best detection rate for the DoS attack, and Multi-classifier [13] showed good detection rate for the Probe and U2R attacks. Works by Xuren et al. [12] provided the best performance for the normal class. Our proposed method demonstrated a better detection rate for R2L attacks and provided comparable performance for Probe and U2R attacks.

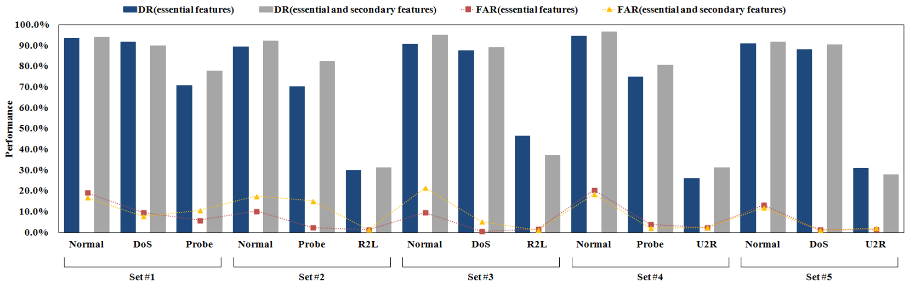


Fig. 4. Comparison of performance results: DR and FAR

Table 1. Comparison of performance results with other works

Method	Normal	DoS	Probe	R2L	U2R
Proposed method with the essential features	0.917	0.890	0.719	0.381	0.284
Proposed method with the essential and secondary features	0.937	0.897	0.802	0.342	0.295
SVM with BIRCH clustering (Horng et al., 2011)	0.993	0.995	0.975	0.288	0.197
ESC-IDS (Toosi et al., 2007)	0.982	0.995	0.841	0.315	0.141
Association rule (Xuren et al., 2006)	0.995	0.968	0.749	0.079	0.038
Multi-classifier (Sabhnani et al., 2003)	n/r	0.973	0.887	0.096	0.298

5 Conclusion

In this paper, an HFR method that combines hierarchical clustering and factor analysis was introduced; an anomaly detection approach based on an MLR was presented. Experimental results show that the proposed system could achieve a significant detection performance by using only a half percent of the comparison feature-set and 39.0% as compared with the full features. Our method also improved detection times by 23.8% compared to those including the secondary level features. Therefore, it can be concluded that our approach can efficiently reduce the features that are redundant or that hinder the process of detecting intrusions. The proposed method enabled reinforcing detection efficiency by the detection of multi-category attacks as an outcome. Future research will include the integration of various probabilistic techniques to achieve better detection performance and the accuracy of predictions.

References

1. Kabiri, P., Ghorbani, A.A.: Research on Intrusion Detection and Response: a Survey. Int. J. Netw. Sec. 1, 84–102 (2005)
2. Lazarevic, A., Ozgur, A., Ertoz, L., Srivastava, J., Kumar, V.: A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection. In: SIAM International Conference (2003)
3. Leung, K., Leckie, C.: Unsupervised Anomaly Detection in Network Intrusion Detection. In: Australasian Computer Science Conference (2005)
4. Chan, P.K., Mahoney, M.V., Arshad, M.H.: Learning Rules and Clusters for Anomaly Detection in Network Traffic. In: Managing Cyber Threats: Issues, Approaches and Challenges, pp. 81–99. Springer (2005)
5. Valdes, A., Skinner, K.: Adaptive Model-based Monitoring for Cyber Attack Detection. In: Recent Advances in Intrusion Detection Toulouse, pp. 80–92 (2000)
6. Xu, J., Shelton, C.R.: Intrusion Detection using Continuous Time Bayesian Networks. J. Art. Int. Res. 39, 745–774 (2010)

7. Huang, L., Nguyen, X., Garofalakis, M., Jordan, M.I., Joseph, A., Taft, N.: In-Network PCA and Anomaly Detection. In: Neural Information Processing Systems, pp. 617–624 (2006)
8. Toosi, A.N., Kahani, M.: A New Approach to Intrusion Detection based on an Evolutionary Soft Computing Model using Neuro-Fuzzy Classifiers. Com. Comm. 30, 2201–2212 (2007)
9. McFadden, D.: Conditional LogitAnalysis of Qualitative Choice Behavior. Frontiers in Econometrics, 105–142 (1974)
10. Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A.: A Detailed Analysis of the KDD CUP 99 Data Set. In: IEEE Symposium on Computational Intelligence for Security and Defense Applications (2009)
11. Horng, S.J., Su, M.Y., Chen, Y.H., Kao, T.W., Chen, R.J., Lai, J.L., Perkasa, C.D.: A Novel Intrusion Detection System based on Hierarchical Clustering and Support Vector Machines. Exp. Sys. W. Appl. 38, 306–313 (2011)
12. Xuren, W., Famei, H., Rongsheng, X.: Modeling Intrusion Detection System by Discovering Association Rule in Rough Set Theory Framework. In: International Conference on Computational Intelligence for Modelling Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (2006)
13. Sabhnani, M.R., Serpen, G.: Application of Machine Learning Algorithms to KDD Intrusion Detection Dataset within Misuse Detection Context. In: International Conference on Machine Learning: Models, Technologies, and Applications, pp. 209–215 (2003)