ISSN 2229-5518

Anomaly Detection through NN Hybrid Learning with Data Transformation Analysis

Saima Munawar, Mariam Nosheen and Dr.Haroon Atique Babri

Abstract— Intrusion detection system is a vital part of computer security system commonly used for precaution and detection.It is built for classifier or descriptive or predictive model to proficient classification of normal behavior from abnormal behavior of IP packets. This paper presents the solution regarding proper data transformation methods handling and importance of data analysis of complete data set which is apply on hybrid neural network approaches for used to cluster and classify normal and abnormal behavior to improve the accuracy of network based anomaly detection classifier. Because neural network classes only require the numerical form of data but IP connections or packets of network have some symbolic features which are difficult to handle without the proper data transformation analysis. For this reason, it got non redundant new NSL KDD CUP data set. The experimental results show that indicator variable is more effective as compared to the both conditional probabilities and arbitrary assignment method from measurement of accuracy and balance error rate.

Index Terms — ANN, Anomaly Detection, Self Organizing Map, Backpropagation network, Indicator variables, Conditional probability

1 INTRODUCTION

In computer security, network administrators always suggest prevented action for better cure of any system. Intrusion Detection Systems (IDS) are classified in to three categories which are host-based, network-based and vulnerability-assessment [1].Signature based detection and anomaly detection model are two basic models of intrusion detection. In signature based, it is only used to detect attack through known intrusions and it cannot be detected novel behavior. It is specially used in commercial tools and it has to update new attacks in database.The anomaly intrusion detection can be resolved these limitation of signature based and used to detect new attack via searching abnormality [2], [3]. Anomaly detection issues have numerous possibilities that are yet unexplored [4]. Network and computer security is significant issues of every security demanded organization. Prevention, detection and response are three basic foundation of network security.For this purpose many researchers emphasizes on preventive action over the detection and response [5]. For increasing the demand of network security, many devices like firewall and intrusion detection used to contol the abnormal packets accesibility.Basically abnomal packets violate the internet protocol standards and these packets is used to crash the systems [6].So this reason better intrusion detection devices are building for prevention and accurate detection of normal and abnormal packets and to reduce the false alarm rate. IDS are basically devoted to fulfill this purpose to monitor the system intelligently. As far as the access control points is concerned ,firewall is good but it is not designed to prevent action against intrusions that's why most security experts emphasizes the IDS which is located before and after the firewall [7], [8].Many researchers have been improving intrusion detection systems through different research areas such as statistics, machine learning, data mining, information theory and spectral theory[2], [3] [4].The purpose of this research is to provide the hybrid learning of artificial neural network base design approach for anomaly intrusion detection classifier system.There is unable to directly handle the symbolic features of IP data set so that It is considered that there are two data transformation methods indicator variable and conditional probabilities which are effective to improve the classifier performance, it is processed through hybrid technique self organizing map and backpropagation of neural network.The data transformation is processed on selecyive nine features of IP NSL data set.It is prepared for anomaly detection classifier which is used for LAN security.

Five sections are presented in this research. Section 2 is background literature of the related research processes. Section 3 provides the detail analysis of proposed research methodology, algorithms of SOM and BPN and their training and testing results are discussed. Section 4 provides detail analysis of experimental results of the research and comparison between

- ***Saima Munawar*** *is with Computer Science department as research fellow at LCWU, Lahore, Pakistan. She is currently working as faculty member in VU, Lahore, Pakistan (e-mail: saima.munawar@vu.edu.pk).*
- ***Mariam Nosheen*** *is with the Computer Science Department as Assistant Professor in LCWU, Lahore, Pakistan (e-mail: m_sufyan2000@yahoo.com).*
- ***Dr.Haroon Atique Babri*** *is with the Electrical Engineering Department as Professor in UET Lahore, Pakistan (e-mail: babri@uet.edu.pk).*

two methods effect the performance of classifier. In last, section 5 presents conclusion and discussed the future direction of this domain.

2.Related study

2.1 Hybrid learning use in misuse and anomaly detection

Hybrid approaches have been used to resolve the anomaly intrusion detection problems. Hamdan et.al [9] comparison four techniques of supervised learning of support vector machine and neural network self organizing map and fuzzy logic of unsupervised learning techniques. It is only proposed descriptions of theses techniques but did not include the methodology and numerical analysis of all these applied techniques. Other approach artificial immune system is used for detection and self organizing is used for classification. It is emphasized the higher level information output rather than the low level for more beneficial to security analyst to analyzing reports. The KDD CUP 1999 data set is used as input for specially focused on two types of attacks which is denial-of-service and user-to-root attacks [10].M.bahrololum et .al [11] presented the design approach and it would be used further explanation in future enhancement. It described introduction of SOM and backpropagation algorithm, KDDCUP data set features, training and testing data, experimenting table view. But besides all of these it did not mentioned how to arrange and used this data set in to which software, how to implement experiment, how to apply these techniques on data set and what methods used to evaluate result. It only provided the proposal and discussed some design issue with flow diagrams. Hayoung et al [12] proposed the new labeling methods apply for this domain but in real time system detection, if no correlation build how to detect the normal or anomaly data set but labeling is supervised learning ,again a huge analysis will require for the correlation between the features. But it did not provided the labeling time only described the detection time but in real time system total time is require for the completion of all processes.

2.2 Analysis and Data Transformation Processes

The data analysis and preprocessing is core part of the artificial neural network architecture for processing and analysis of accurate result. Anomaly detection has been paying attention of many researchers during the last decade. Due to this reason many researcher not only considered the new algorithms but also taking analysis of data set used for training and testing classifier. The KDD CUP 99 data set is mostly used for intrusion detection problems. It has 41 features. There are three basic features which are individual TCP connections feature, content features, and traffic features which include 7 symbolic and 34 continuous attributes [13]. Tavallaee et.al presented the detail and critical review of KDDCUP99 data set. It is discussed the problems in KDD CUP99 data set and resolved two issue of KDD CUP 99 data set which affects the performance and poor evaluation in anomaly detection approaches. It proposed new data set NSL-KDD, which include selected records and remove redundancy of records of KDD CUP 99.The form of this data set is ARFF (attribute relation file format).The authors claimed that this data set will help researcher for solving anomaly detection problems [14].Preprocessing apply before processing of neural networks algorithms because these algorithms require the quantitative data instead of qualitative information.The most commonly conversion method used is arbitrary assignment but criticizing of this method, three other approaches is using for machine learning algorithms. E.Hernandez et.al presented three methods for symbolic features conversion apply on KDD CUP data set. It described all these techniques in detail and also described the comparison of these techniques have been applied on different feed forward neural network and support vector machine. They claimed that these three conversion methods improve the prediction ability of the classifier. These methods are using for preprocessing (symbolic attributes convert in to numeric form) which is indicator variable, conditional probabilities and SSV (Separability split value) criterion based method [15].

3. Proposed experimental methodology

This section is divided into theses main processes which are data Analysis, preprocessing, modeling of clustering and classification and performance evaluation.

3.1 Data Analysis

NSL KDD CUP data set is reasonable and improves the evaluation.This data set is offline and it is provided for anomaly detection classification research to better evaluation of classifier.It also gives the consistent and getting more comparable results [14], [17], [18].

3.2 Feature selection

It is difficult to select the important feature for detecting and classification between normal packets and attacks. More research work is doing for selection of feature on anomaly detection problems.The basic question is how many types of features are selective for improving the classification rate and to relate which types of attacks. In this research, first basic 9 attributes of individual tcp connections are used. It consist of duration, types of protocol, services, source bytes, destination bytes, flag, land, wrong fragment and urgent. These features have 3 symbolic and 5 continues attributes. The protocol and services are most important features to detect the attacks [13], [14].The main purpose to select these features because it has maximum number of symbolic features instead of others for handling symbolic features preprocessing.

3.3 Preprocessing

The given input data set has symbolic and continuous attributes. These data set need to be converted in to numerical form for processing on neural network algorithms. Researchers are finding best data transformation techniques applying on selected features for improving the performance of classifiers. The main purpose is to show how different preprocessing methods affect the accuracy of different tasks of machine learning simulation.Besides the modification of algorithm, it also important to consider data transformation

and feature selection methods according to the demand of any machine learning and training.The details of data transformation methods are used in this research which is given below.

3.3.1 Indicator Variables Method

The basic procedure of indicator variables is that, 1 indicates the occurrence of categories of features and 0 indicates its nonoccurrence of categories of features. The nominal feature X represents the features with N distinct categories, a set of N indicator variables can be generated it depend on the categories exist in the attributes of data set [19].We have successfully processed first indicator variable method on the basic symbolic features of data set but after preprocessing the dimensionality of attributes is increased. In the research nine basic features are used but after apply this method it grows and it is change in to 122 attributes. This problem is reduced through clustering techniques apply on large number of categories of attributes such as protocol, services and flag attributes [15].

3.3.2 Conditional probability

The second method is to convert a symbolic feature with an array of conditional probability for obtaining each class given the attribute that has particular symbolic value. In this case each symbolic value x_k of a feature 'a' may be replaced by the following N-dimensional vector of conditional probabilities [15],[20].N is the number of classes of the training set and T is number of categories of the symbolic value of xk.We have successfully applied this method on symbolic features of data set.This method is given below

$$\sum_{k=1}^{T} (CP(1 \mid a = x_k), CP(2 \mid a = x_k), CP(3 \mid a = x_k), \ldots CP(N \mid a = x_k))$$

3.4 Modeling

The hybrid model of neural network which is the combination of self organizing map and two layers feed forward BP network. This model detail is discussed below

3.4.1 SOM Clustering

SOM is used for clustering and also visualizing high dimension data set.It is basically used to find the similarity map of the input data. There are usually three types of topologies used in the SOM mapping, which is rectangular, hexagonal and random mapping grid. The distance is found from usually four different ways like Euclidean, box, link and manhattan distance [16]. The basic SOM algorithm is summarized below

1. Firstly set the topology of neurons for SOM mapping.
2. Initialize the input weight 'wij' from random selection of weight vector for neuron ' ij ' , since i = 1, ..., m , j = 1,…,n
3. The input 'x' is also select randomly from given data set but it takes input as different dimension of features set so firstly transpose the original data.
4. The SOM basic purpose is to find winning neuron , it is also called hit ' h ' which is determine through different distance function like it can be shown in equation 1 which is build by Euclidean distance.

$$\|w_h - x\| = \min \| w_j - x \| \ldots\ldots 1$$

Table 3.1: The training network of SOM clustering

NN Architecture	Training function	Preprocessing Methods	Neuron Dimension	Epoch	Completion Time
SOM	Batch unsupervised Weight/bias	IV	10 ×10	500	0:11:47
		CP	10 ×10	500	0:01:31

5. All topology neurons ' ij ' in the neighborhood of winning neuron "h" which is updated all weight vectors by usually used gaussian function apply between the two neurons distance in output layer. This function φ(j,h) represents the connection of neuron j and h which is closely related to each other. It is determine by eq 2

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t) \cdot \varphi(j,h)(t) \cdot \{x_i - w_{ij}(t)\} \ldots\ldots 2$$

η represents learning rate at the epoch t.

6. Repeat the steps 2 to 4 until its satisfied the convergence criteria.This algorithm is operated for (m*n) dimension of topology of neurons with respect to input [21], [22].

3.4.1.2 SOM Training and Testing

In Section 3.3 two preprocessing methods are prepared from selective amount of training and testing data sets and it is further used for clustering.The training and testing has been performed on Matlab 7.9 software with visualization of new self organizing map plotting [16].The original input is firstly transpose because SOM algorithm takes the features as rows and takes samples as columns. After the preparation of SOM input, the plotting of both two methods IV and CP.In the experiment, hexagonal topology map for output layer has been settled with 10 × 10 dimension of 100 neurons. It shows in fig 3.1. The network of SOM is created with topology and input of the data set.

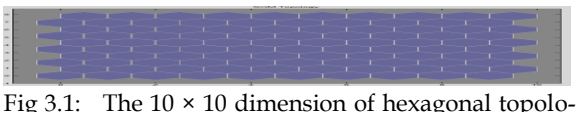

Fig 3.1: The 10 × 10 dimension of hexagonal topolo-

Firstly the weight is initialized. The batch learning algorithm is faster than incremental learning so this reason the batch unsupervised weight training function is used in SOM learning and their parameter of initial neighborhood size is 3 and the steps of ordering phase is 100.It is trained with 500 epochs or cycles. The training information of both IV and CP methods with their completion time are given in Table 3.1. To get the bases of average features of each neuron through network weight layers. Simulate the network with trained network to get an output frequency of each neuron. Those neurons which have maximum frequency of connection records, it is called hit of the map, the 100th location of neuron topology gives the hit of SOM through the bases of weight center. The mapping of this winning neuron and

neighbor neurons with frequency of records of both methods shows in fig 3.2 and fig 3.3.In this research only include the winning neuron analysis which is in 100th location. It also got average weights of each features through network weight layers.The building of 100 neurons is based on these weight average clusters and its plotting also shows the input and weight centers for determine the clusters in fig 3.4 and fig 3.5, the circle shows the weight center and + sign shows the input. After getting output of training SOM, simulate the testing or unseen output data is put on trained network. Similar to training, the 100th topology of neuron gives the winning neuron output of testing. After getting output of training SOM, simulate the testing or unseen output data is put on trained network. After get index of winning neurons of training and testing data set. It is related to both different methods for the classification of normal and anomaly data set.

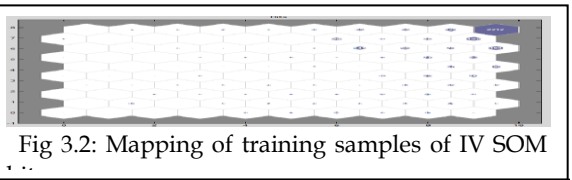

Fig 3.2: Mapping of training samples of IV SOM

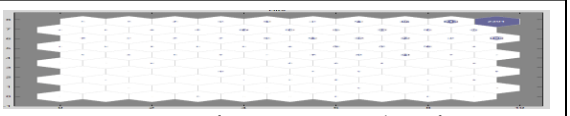

Fig 3.3: Mapping of training samples of CP SOM

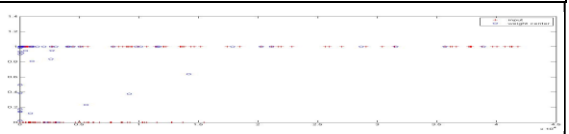

Fig 3.4 Plotting of input and weight center of IV

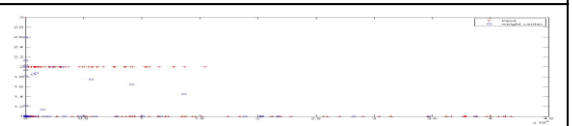

Fig 3.5 Plotting of input and weight center of CP

3.4.2 Backpropagation classification

In this research two layer feed forward backpropagation is used for classification of normal and anomaly behavior. This network architecture consists of input, weight, bias, hidden layer and output layer. A sigmoid transfer function is used for hidden layer and output layer is managed by linear function. The different types of gradient descent training algorithms are used for updating the weight and biases.The BPN two layer feedforward algorithm is summarized below

1. Firstly in the network, randomly initialized all weights
2. The samples of data set provide to network in the pairs of vectors 'x' represents the input and 'y' represents target.

$$\{(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_s, y_s)\} \ldots\ldots 1$$

3. The computation of network learning start from put examples in the network and it compute all neurons output until get output of network 'O'
4. The algorithm iteratively reduces the error through gradient descent function. For using δ^o_{su} to compute the error between's ' represents samples across U represents all output layer.

$$\delta^o_{su} = (y_u - o_u) f'(net_u^o) \ldots\ldots\ldots 2$$

The representation of basic architecture of feed forward BPN in Matlab 7.9 [16].

3.4.2.1 BPN Training and Testing

The training output of both methods SOM winning neurons are used as input for BPN classification. The BPN network is created with 20 hidden neurons. Firstly input is partitioned in to three ratios for better classification, 90% is used for training, 5% is used for validation and 5% is used for testing. The testing output of both IV and CP method of SOM winning neurons are used as unseen data for check the performance of classification. The training is used scaled conjugate gradient backpropagation function to apply the network. Simulate the network with trained network to get an output of normal and anomaly classes to each samples of winning neuron. The classification is completed after 28 epochs. After completing the training of classification, the output generate mean square error which indicate average square difference between output and target and percent error indicates the fraction of error which are not classified. The summary of all classification training parameter output is illustrated in Table

Table 3.2: The output of classification training parameters

Preprocessing Methods	Data Division	MSE	%E	Gradient	Overall performance MSE	Epoch Completion	Time Completion
IV	Training	0.068	7.19E+00	0.0048	0.04903	28	0:00:03
	Validation	0.066	7.21E+00				
	Testing	0.02	1.80E+00				
	Unseen data	0.042	4.25E+00				
CP	Training	0.079	8.59E+00	0.0054	0.05987	28	0:00:04
	Validation	0.095	1.04E+01				
	Testing	0.052	5.22E+00				
	Unseen data	0.013	1.47E+01				

3.2

4. Experimental Result

As described in the section 1, this work is not compared the performance measure of clustering and classification algorithms but it examined the data transformation methods apply before these algorithms and their effectiveness over the performance of classifier or model. The first experiment has performed by SOM clustering which involved the training and testing data set of selective featuresof NSL KDD CUP data set.This experiment output is given to the two layer feed forward BP network for the classification to determine normal and anomaly classes exist in each wining neurons. The result shows these winning neuron are related to anomaly cluster which has maximum number of samples relate to anomaly class and minimum relate to normal classes. For the IV methods, the total number of samples exist in training winning neuron is 2212 samples which is determine by SOM.The 2059 samples relate to anomaly class and 153 samples relate to normal class, it is determine by BPN classification. This means, this winning neuron is related to anomaly cluster based neuron. By examination their performances we have simulate unseen data on it. The total number of testing winning neuron has 1858 samples and through classification determination it is also related to anomaly class which has 1779 examples relate to anomaly class and 79 relate to normal class. After the measurement of their accuracies, it shows the training winning neuron is 93.1 % accuracy and testing neuron is 95.7 % accuracy. Similar to this procedure applies on the CP method which has 2301 samples of training winning neuron. It is determined that 2105 relate to anomaly class and 196 relate to normal class and to check its performance by unseen data simulate on it. The total number of samples exist in testing winning neuron is 1751 samples. From classification to determine 1494 relate to anomaly class and 257 relate to normal class. This gives 91.5 % accuracy of training winning neuron and 85.3 % accuracy of testing neuron. The both methods are maximum number of correctly classified and minimum misclassification occurred. Through unseen data set examination, it is also related to anomaly cluster neuron. It is illustrated in table 4.1. The comparison of preprocessing methods show that arbitrary assignment method which is applied on multilayer feed forward network, it accuracy was 80.88%.In this research, the accuracy of IV and CP are 95.7% and 85.3% which is applied on unseen or testing data it can be shown in Table 4.2. The indicator variable (IV) shows better performance than the conditional probabilities (CP) and previously used arbitrary assignment (AS) conversion method using neural network approaches for network anomaly based classifier. The results of both are also shows in the graph of ROC curve which show the true positive and false positive rate of winning neuron. It can be shown in fig 4.1 and 4.2. The analysis of both comparison methods of winning neurons show that indicator variable method gives the better performance as compare to conditional probability method. The overall performance is measured through the determination of balance error rate (BER) on unseen data.which is applied on all neurons examination but it is estimated measurement of all neurons. Through this examination we can compare which has generated less error. In last, the determinations of BER shows that the applied indicator variable method has 0.3057 error rate which is less error rate as compare to the conditional probabilities method which has 0.4459 error rate. It is illustrated in Table 4.3.

Table 4.1: Results of winning neuron confusion matrix

Preprocessing methods	Data Division	Normal Class	Anomaly (Attacks) Class	Correctly Classified %	Incorrectly Classified %	Overall Accuracy %
IV	Training	143	1847	92.8	7.2	93.1
	Validation	8	103	92.8	7.2	
	Testing	2	109	98.2	1.8	
	Unseen data	79	1779	95.7	4.3	95.7
CP	Training	178	1893	91.4	8.6	91.5
	Validation	12	103	89.6	10.4	
	Testing	6	109	94.8	5.2	
	Unseen data	257	1494	85.3	14.7	85.3

Table 4.2: Comparison of Preprocessing Methods

Testing data	Preprocessing Methods		
	AS	IV	CP
Overall accuracy%	80.88	95.7	85.3

Table 4.3: Balance error rate of both methods

Preprocessing methods	Data	Balance error rate (BER)
IV	Unseen data set	0.3057
CP	Unseen data set	0.4459

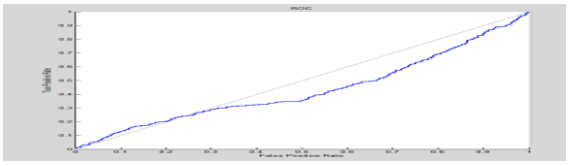

Fig 4.2: Unseen data of ROC curve of CP

5. Conclusion and Future aspiration

The main conclusion has been achieved from experiments that indicatior variables and conditional probability method are correctly detected the anomlay based winning neuron by SOM clustering.Basically the winning neuron output of SOM is used as a input for BPN classification in this research.The BPN classify the normal and anomaly classes of winning neuron. The testing accuracy of IV winning neuron is 95.7 % which is better than 85.3% acuracy of CP.The 10.4% difference of accuracy shows from this camparison. The comparison of two models are also measured by BER which is applied on unseen data of hit neurons.This measurement shows the indicatior variable has generated less error rate as compared to conditional probabilities performance.The analysis of all 100 neurons is needed which exists number of anomaly and normal neuron. It should be done only by unsupervised way to detect behaviour. In this research, it is also provided the bases of cluster building.This gives the average of each features of the samples in the form of weight center.It is suggested that it is beneficial to find the normal and anomaly behaviour through any condition developing based or rule based algorithm like genetic algorithm.

REFERENCES

[1] Jean-Philippe. (2001). Retrieved from http://www.sans.org/reading_room/whitepapers/detection/application-neural-networks-intrusion-detection_336

[2] Chen, C.-M., Chen, Y.-L., & Lin, H.-C. (2010). An efficient network intrusion detection. Computer Communications, 33, 477–484.

[3] Bahrololum, M., & Khaleghi, M. (2008). Anomaly Intrusion Detection System Using Hierarchical Gaussian Mixture Model. IJCSNS International Journal of Computer Science and Network Security, 8.

[4] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly Detection: A Survey. University of Minnesota. ACM Computing Surveys.

[5] Stoneburner, G. (2001). Underlying Technical Models for Information Technology Security. National Institute of Standards and Technology. WASHINGTON: NIST Special Publication 800-33.

[6] Frederick, K. K. (2000). Abnormal IP Packets. Retrievedfromhttp://www.symantec.com/connect/articles/abnormal-ip-packets

[7] Dupuis, C. (2002). Intrusion Detection Systems (Ids). Retrievedfromhttp://www.cccure.org/Documents/IDS/IDS_2002.PPT

[8] Intrusion Prevention Systems (IPS). (2008). Retrieved from http://nsslabs.com/white-papers/intrusion-prevention-systems-ips.html

[9] Alanazi, H. .., Noor, R. M., Zaidan, B., & Zaidan, A. (2010). Intrusion Detection System: Overview. Journal of computing, 2(2).

[10] Powers, S. T., & He, J. (2008). A hybrid artificial immune system and Self Organising Map for network intrusion detection. Information Sciences, 178 , 3024–3042

[11] Bahrololum, M., Salahi, E., & Khaleghi, M. (2009). Anomaly Intrusion Detection Design Using Hybrid of Unsupervised And Supervised Neural Network. International Journal of Computer Networks & Communications, 1(2).

[12] Oh, H., Doh, I., & Chae, K. (2009). Attack classification based on data mining technique and Its application for reliable medical sensor communication. International Journal of Computer Science and Applications, 6(2), 20 – 32.

[13] KDD Cup 1999 Data. (1999). Retrieved from http://kdd.ics.uci.edu/databases/kddcup 99/kddcup99.html.

[14] Tavallaee, M., Bagheri, E., Lu, W., & Ghorban, A. A. (2009). A Detailed Analysis of the KDD CUP 99 Data Set. IEEE Symposium on computational Intelligence in Security and Defence Application.

[15] Hernández-Pereira, E., Suárez-Romero, J. A., Fontenla-Romero, O., & Alonso-Betanzos, A. (2009). Conversion methods for symbolic features: A comparison applied to an intrusion detection problem. Expert Systems with Applications, 36, 10612–10617.

[16] Demuth, H., Beale, M., & Hagan, M. (2010). Neural Network Toolbox™ 6. Retrieved from www.mathworks.com/access/helpdesk/help/pdf_doc/nnet/nnet.pdf

[17] The 16 ACM SIGKDD conference on knowledge discovery and data mining. (2010). Retrieved from http://www.kdd.org/kdd2010/tutorials.shtml

[18] The NSL-KDD Data Set.(2009).Retrieved from http://nsl.cs.unb.ca/NSL-KDD/

[19] Shyu, M.-L., Sarinnapakorn, K., Kuruppu-Appuhamilage, I., Chen, S.-C., Chang, L., & Goldring, T. (2005). Handling Nominal Features in Anomaly Intrusion Detection Problems. 15th international workshop on research issues in data engineering:strean data mining and applications.

[20] Duch, W., Grudzinski, K., & Stawski, G. (2000). Symbolic Features in Neural Networks. Torun, Poland.

[21] Pomplun, M. (2006). Artificial Neural Network Paradigms. Retrievedfrom http://www.cs.umb.edu/~marc/www.scs.ipm.ac.ir/seminars/Lecture/.../mark%20pomplan/marc.../talk8.ppt

[22] Gunes Kayacik, H., Zincir-Heywood, A. N., & Heywood, M. I. (2007). A hierarchical SOM-based intrusion detection system. Engineering Applications of Artificial Intelligence, 20 , 439–451.