IJE TRANSACTIONS A: Basics Vol. 30, No. 10, (October 2017) 1494-1502

International Journal of Engineering

Journal Homepage: www.ije.ir

A Novel Intrusion Detection Systems based on Genetic Algorithms-suggested Features by the Means of Different Permutations of Labels' Orders

J. Ghasemi*[a], J. Esmaily[b]

[a] *Faculty of Engineering & Technology, University of Mazandaran, Babolsar, Iran*
[b] *Shahid Rajaee Teacher Training University, Tehran, Iran*

PAPER INFO

Paper history:
Received 09 April 2017
Received in revised form 03 May 2017
Accepted 07 July 2017

Keywords:
Intrusion Detection Systems
Feature Selection
Genetic Algorithms

ABSTRACT

Intrusion detection systems (IDS) by exploiting Machine learning techniques are able to diagnose attack traffics behaviors. Because of relatively large numbers of features in IDS standard benchmark dataset, like KDD CUP 99 and NSL_KDD, features selection methods play an important role. Optimization algorithms like Genetic algorithms (GA) are capable of finding near-optimum combination of the features intended for construction of the final model. This paper proposes an innovative method called chain method, for evaluation of the given test record. The main intuition of our method is to concentrate merely on one attack type at every stage. In the beginning, datasets with the proposed features by GA based on different labels will be assembled. Based on a specific sequence– which is found on different permutation of four existed labels- the test record will be entered the chain module. If the first stage –which is correlated to the input sequence-, is able to diagnose the first label, the final output has been indicated. If is not, the records will pass through the next stage until the final output be obtained. Simulations on proposed chain method, illustrate this technique is able to outperform other conventional methods especially in R2L and U2R detection with the accuracy of 98.83% and 98.88% respectively.

doi: *10.5829/ije.2017.30.10a.10*

1. INTRODUCTION

There are no doubts, internet attacks are truly precarious and must be taken seriously. Intrusion Detection Systems (IDS) are one of the most applicable and useful methods to diagnose attack traffics patterns. These systems utilize machine learning methods to construct a smart model for attacks recognition. Researchers generally seek to build their models on a reliable and studied dataset. In the IDS discussions, KDD CUP 99[2] is one of the most famous and studied datasets ever. However, it is recommended to benefit from more recent datasets which are able to address more fresh and advanced attack types. A more recent version of KDD CUP 99 dataset is NSL_KDD[3] [1] which are structurally similar to well-known KDD dataset but of course, contain more reliable and applicable attack varieties. Moreover, advances in computer networks and implementation of novel topologies for packages transfer advocate different pattern and behavior in new network traffic. Hence, shifting to a new dataset is crucial for real world applications.

Both datasets like other standard datasets in this scope contains numerous features. The features which are able to divide the space into more discriminative subspace, are more encouraged. Curse of dimensionality suggests the results of choosing more features will not be always promising [2, 3]. Used datasets in this paper convey this consequence as Erfani et al. [4]. In this way, features selection method is able to counter this effect as one of faced problem in IDS discussions.

*Corresponding Author's Email: *j.ghasemi@umz.ac.ir* (J. Ghasemi)

[2] kdd.ics.uci.edu/databases/kddcup99/kddcup99.html/

[3] http://www.unb.ca/research/iscx/dataset/iscx-NSL-KDD-dataset.html

Please cite this article as: J. Ghasemi, J. Esmaily, A Novel Intrusion Detection Systems based on Genetic Algorithms-suggested Features by the Means of Different Permutations of Labels' Orders, International Journal of Engineering (IJE), TRANSACTIONS A: Basics Vol. 30, No. 10, (October 2017) 1494-1502

2. RELATED WORKS

KDD CUP 99 is applied as the main benchmark dataset since of its appearance and has been counted as the major dataset for IDS researches [5, 6]. An improved version of well-known and well-studied KDD CUP 99 emerged and called NSL_KDD [1, 7, 8]. By exploiting optimization algorithms like GA, the near optimum combination of features is obtainable. In classification phase, DT [9, 10], MLP [11, 12], KNN [13, 14] and SVM [15, 16] can be measured as the most applicable and popular classification algorithms.

Using classification algorithm alongside of evolutionary algorithms such as GAs [17] becomes more popular in recent years. Methods like GA are more helpful especially when the size of space becomes much larger [18]. In literature [19] with the combination of statistical concepts and SOM algorithm, authors are able to reach an efficient model on NSL_KDD dataset. They used Principle Component Analysis (PCA) and Fisher Discriminant Ratio (FDR) for feature selection and noise removal phase, respectively. For diagnosing traffic types, the SOM algorithm has been applied. In reference [20], the concept of multi-objective approaches for feature selection and its application in Growing Hierarchical Self-Organizing Maps (GHSOMs) has been utilized. Moreover, with multi-objective approach, they are able to distinguish between attack and normal and of course between attack types. Authors use NSL_KDD as their benchmark. In Ref. [6], authors used a hybrid KPCA_SVM_GA method for attack detection. They used KPCA for feature reduction, SVM in classification phase and GA for setting punishment factor of a parameter of C in kernel function of SVM. Simulations on KDD CUP 99 confirmed their model is efficient in IDS criteria evaluations. In Ref. [21], with an innovative idea, the authors present a biologically-inspired computational approach learn signatures for network traffics using a supervised learning classifier system. Minimization of the overlaps and conflicts between signatures by new generalization operator represents their main approach. They used KDD CUP 99 dataset as their evaluation benchmark as suggestion of the effectiveness of their model. In Ref. [22], the authors used a fuzzy rule-based system which can act as a genetic feature selection for finding optimal feature combination. Their simulations on KDD CUP 99 dataset show improvement in IDS criteria in comparison with general simulation. In Ref. [23], cuttlefish algorithm (CFA), an optimization algorithm uses as a search strategy to extract an optimal subset of features and the decision tree (DT) classifier for the classification phase. KDD CUP 99 has been put in the consideration for built model evaluations. In comparison with all features, obtained results show better performance. Last four spoken studies used old KDD CUP 99 dataset and as discussed before, it is not able to monitor novel attack types. These studies performance on new datasets like NSL_KDD is not yet determined. Hence, their reliability and performance in real work environment is not completely trusted. In Ref. [24] based on KDD CUP 99, they used a framework which combines multiple classifier outputs in order to enhance performance. They introduced the novel Multiple Adaptive Reduced Kernel Extreme Learning Machine (MARK-ELM) which combines Multiple Kernel Boosting with the Multiple Classification Reduced Kernel ELM. In Ref. [25] authors exploit the adapted chaos concept in their proposed time-varying chaos particle swarm optimization (TVCPSO) method to carry out parameter setting and feature selection for multiple criteria linear programming (MCLP) and support vector machine (SVM). NSL_KDD as a more reliable version of KDD CUP 99 has been chosen for their evaluations. In Ref. [26], authors use an outlier based system based on identifying relevant subset of features by mutual information and generalized entropy-based feature selection algorithms. A tree-based clustering technique also has been used for ranking outlier and finding anomalies. They used NSL_KDD dataset for their complex method evaluations. Last two discussed studies use NSL_KDD dataset for their simulations which make the outcomes more consistent and applicable. Last three papers have been compared with our proposed method to demonstrate proposed method's ability versus recent studies. As it has been illustrated further in this paper, our proposed method is able to outperform the results of these three papers.

For IDS studies, algorithms like GA is commonly used as a feature extraction tool [26]. GA is a population-based optimization algorithm which tries to attain optimum solution by its efficient operators like crossover and mutation. In this paper, GA attempts to examine the different combination of features and reach to near optimum accuracy or detection rate for a specific label. It will return a combination of features which are able to achieve fine performance based on four attack types in the dataset. Different permutation of four attack types in the KDD and NSL dataset will generate 24 different sequences as proposed method input and consequently 24 different performances. As it is publicized further in this article, focusing merely on a specific feature in every stage turns to be a very effective method for attack detections.

The contents of this paper will be as follow: In the following section, the proposed method will be explained. In the next section, the results of our method and the performance of proposed method versus other methods will be brought. Best sequences for proposed method will be put in the discussion in the subsequent section. Finally, the conclusion of this study will be conveyed.

3. PROPOSED METHOD

If there will be a mechanism which is able to focus on every attack type and not confused with other labels, the improved results could be expected. Proposed method attempts to work on GA suggested features for a label in order to diagnose that specific label. The system acts as though there is only this attack type in the original dataset by considering all other labels as one class. Trained classifier by modified 2-labeled dataset with GA suggested features for the first label determines whether the entered test record is related to the first label or not. In the case of no relation, the same scenario will be conveyed in the second stage. This mechanism similarity to a chain encourages this paper's authors to name this method as CHAIN method. Based on different permutations of attack types, every label will be detected in the corresponding stage. The sequence input in the chain module can make effective differences. In other word, this is really important which label should be concentrated in the first stage and which one should be focused on the last stage of the chain module. Figure 1 visualizes the proposed chain method. In the first phase, the dataset will be prepared. In contrast with NSL, KDD dataset has some redundant records which will be eliminated in data preparation phase. Random sampling in both datasets will generate the final random datasets which will be considered as the main benchmark for simulations. n the next phase, suggested features based on different labels by GA will be extracted. For every label in both KDD and NSL dataset, a modified 2-labeld dataset will be created. For example, 2-labled Dos dataset contains all records in the random dataset but of course with different labeling. In this dataset, all of the Dos records will be considered as one class, and all other labels – three other attack types and normal records- will be considered as the other class. GA will be executed on this dataset and proposed features will be extracted. The GA tries to minimize the value of numbers of misclassified records based on the specific label. For every label, suggested features will be extracted (see appendix 1). In the table in appendix 1, every feature based on executed GA for specific label have been illustrated. For example, third feature in KDD has been suggested by GA based on Dos and Probe 2-labled dataset. Now, these proposed features help to pursue the next phase.

CHAIN method phase conveys the main idea of this paper. Based on different permutation of attack types as chain method input, the different dataset will be made. For example, suppose a specific sequence: 1,2,3,4. Sequence 1 is corresponding to Dos, 2 to Probe, 3 to R2l and 4 to U2R attack type.

Note this is an example sequence and for complete simulation, all of 24 possible orders will be analyzed. Supposing this example, the first dataset will be generated by GA proposed feature for Dos label. That means there will be a dataset with the same records as random dataset but with proposed features by GA for Dos- not all the features. The trained model via 2-labeld, dataset with proposed features by GA for Dos will decide whether the entered test record is Dos or not. If this is true, the specific test record will be classified as Dos attack and the predicted label has been decided. If it is not, the test record will be entered in the second stage. Before entering the second stage, the records with Dos label will be eliminated from the original random dataset.

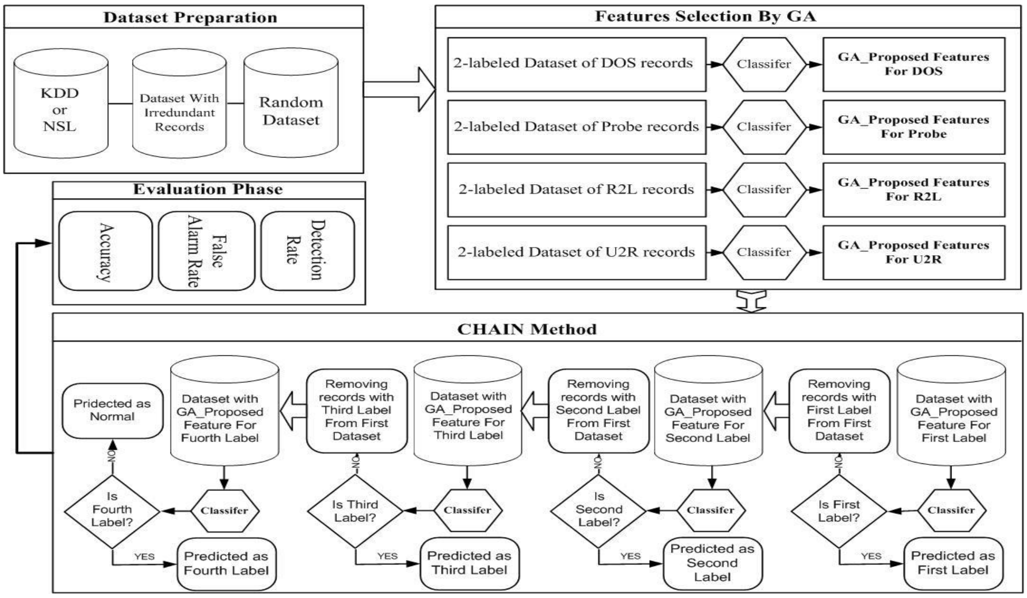
Dataset Preparation
KDD or NSL
Dataset With Irredundant Records
Random Dataset
Features Selection By GA
2-labeled Dataset of DOS records
Classifer
GA_Proposed Features For DOS
2-labeled Dataset of Probe records
Classifer
GA_Proposed Features For Probe
2-labeled Dataset of R2L records
Classifer
GA_Proposed Features For R2L
2-labeled Dataset of U2R records
Classifer
GA_Proposed Features For U2R
Evaluation Phase
Accuracy
False Alarm Rate
Detection Rate
CHAIN Method
Pridected as Normal
Dataset with GA_Proposed Feature For Fuorth Label
Removing records with Third Label From First Dataset
Dataset with GA_Proposed Feature For Third Label
Removing records with Second Label From First Dataset
Dataset with GA_Proposed Feature For Second Label
Removing records with First Label From First Dataset
Dataset with GA_Proposed Feature For First Label
NO
Is Fourth Label?
Classifer
YES
Predicted as Fourth Label
NO
Is Third Label?
Classifer
YES
Predicted as Third Label
NO
Is Second Label?
Classifer
YES
Predicted as Second Label
NO
Is First Label?
Classifer
YES
Predicted as First Label

Figure 1. Block diagram of proposed chain method

The second classifier will be trained on the 2-labeld dataset with GA proposed feature for Probe, which contains all records -except records with Dos label.

If the first stage suggests that the entered record is not Dos, we suggest the records with Dos records should be eliminated to assist the next stage training classifier to consternate on other remaining labels and do not be confused with Dos records to any further extent. This treatment will be followed to the end of the chain. Note at the end of chain method if – for this example- the entered test record do not recognize as U2R by trained classifier on the dataset –which contained no records with class of Dos, Probe and R2L and only includes U2R and Normal records-, the discussed record will be classified as Normal. Based on the predicted outputs by chain module, results will be captured and analyzed in respect of IDS criteria.

In overall, this paper provides an innovative method for considering every attack type with high concentration and moreover a comprehensive and applicable knowledge extraction for input sequences. The accomplished simulations and gathered results will be discussed in the next section.

4. SIMULATION AND RESULTS

Both KDD CUP 99 and NSL_KDD contain numerous records. In this way, dealing with the entire dataset seems a very burdensome task. Therefore, many types of researches like this paper, desire to carry out random sampling from the original dataset as their major dataset [6, 19, 25-29]. In our sampling, redundant records have been removed and the result only contains non-repetitive records [29]. The random dataset generated from KDD has been built on the non-repetitive records. NSL dataset originally doesn't contain any repetitive records; in fact, this characteristic is one of NSL advantages in comparison to KDD. Before analyzing the methods performance and capabilities, the specific criteria for evaluating an IDS need to be explained; Accuracy (A), Detection Rate (DR) and False Alarm Rate (FAR). These measures will be defined as follow:

$$A = (TP+TN) / (TP+TN+FP+FN) \quad (1)$$

$$DR = TP / (TP+FP) \quad (2)$$

$$FAR = FP / (FP+TN) \quad (3)$$

where:

- TP: True Positive: number of Attack records which were classified as Attack
- TN: True Negative: number of Normal records which were classified as Normal
- FP: False Positive: number of Normal records which were classified as Attack
- FN: False Negative: number of Attack records which were classified as Normal

These criteria are the most applicable and the tool for evaluating an IDS could be simply judged for real world applications using these criteria.

In the first simulation, four well-known classification algorithms performances which build their model on the entire features of the dataset have been investigated – we will refer to this simulation as General Simulations. The Table 1 demonstrates the performance of these simulations.

As it has been illustrated in Table 1, in most cases DT have superior performance. This is one of the main reasons why DT has been chosen as the main classifier for further simulations- selection features by GA in fitness function and main classifier in chain module. DT has not the best performance in both datasets only for a few criteria for specific labels. For example, see detection rate for U2R in NSL dataset. R2L and U2R performance in both datasets are undeniably poor.

TABLE 1. Performance of four classification algorithms on the entire features

			DOS	Probe	R2L	U2R	Normal
KDD	DT	ACC.	**97.06**	**84.85**	2.9	**44.28**	**98.63**
	KNN		96.5	78.6	**4**	37.14	97.8
	MLP		82.36	76.05	**4**	35.71	95.16
	SVM		96.86	80	3.2	28.57	97.1
	DT	FA	**1.37**	**1.3**	1.36	1.36	**14.22**
	KNN		2.21	2.2	2.2	2.2	21.85
	MLP		8.17	8.17	8.18	8.02	33.04
	SVM		2.2	**1.3**	**1.1**	**1**	19.7
	DT	DR	**98.61**	**97.64**	**41.42**	**43.05**	**79.26**
	KNN		97.77	95.97	37.73	28.26	69.82
	MLP		90.62	83.23	13.8	7.72	60.46
	SVM		95.7	94.6	26.7	17.5	79
NSL	DT	ACC.	**96.74**	**95.31**	**71**	92.52	**97.45**
	KNN		96.62	92.22	65.46	92.05	83.11
	MLP		85.83	82.27	62.64	87.59	84.31
	SVM		96.4	94	65.66	**93.36**	82.99
	DT	FA	**3.6**	**3.6**	**3.6**	**3.6**	**1.88**
	KNN		4	4	4	4	23.74
	MLP		12.33	12.33	12.33	11.92	24.7
	SVM		4.3	4.3	4.3	4.3	23.53
	DT	DR	**95.29**	**95.13**	**73.72**	41.93	**96.98**
	KNN		94.81	94.23	35.48	38.46	68.27
	MLP		82.43	80.52	16.35	14.66	70.2
	SVM		94.43	94.11	39.43	**49.41**	67.63

For considering all possible permutation of attack types in the chain method, 24 feasible sequences should be entered in the chain module. These 24 sequences have been provided in appendix 2. As it has been shown in the first row, 1 would refer to Dos, 2 to probe, 3 to R2L and finally 4 to U2R. For example, consider the row number 21; the order of considering label in the chain module would be as: U2R, Probe, Dos, R2l. In this paper, all of the possible sequences have been analyzed based on IDS criteria for every dataset – KDD or NSL. Based on the IDS intention in order to counter a specific attack type, a sequence would be selected. There are sequences – like 7 (2, 1, 3, 4) - which are able to outperform the best performance of four popular classifiers in the Table 1. That means alongside possessing sequences with high-quality performance for a particular attack type, there are sequences which are capable of diagnosing all attack types more efficiently without concerning about specifying the attack types.

In first visualization, average of every attack type's results based on different methods will be compared. For every method and criteria, average of five existed label has been calculated. Figure 2 displays the demonstration. In this figure, performance of different method based on different IDS criteria has been exposed. Note that Figure 2(a) represents the KDD simulations while Figure 2(b) illustrates NSL simulations. DT_All means the simulation of DT on a dataset which works on all 41 features. In both datasets, chain method provides superior outcomes. Calculation of average of five labels (Dos, Probe, R2L, U2R, and Normal) for each method is a reliable way to judge on the comprehensive IDS. Earlier in this paper, results of table 1 have been alarmed as one of the main reason why DT has been selected for GA fitness and chain method classifier. Moreover, above figure provides another reason for this matter. As it has been shown in Figure 2, the proposed method is able to outperform other general method in the case of 3 defined criteria.

The next presentation addresses the best performance of IDS criteria. For every attack type, based on IDS criteria, best results of all method have been captured and suitably compared with proposed chain method. In first figure, detection rate and accuracy will be examined.

In both dataset, for R2L and U2R attack types, chain method improves their performance more significantly. As it has been mentioned before, there are sequences which are able to outperform general simulation outcomes solely (see sequence 7).

FAR of every attack types represented in the Figure 4. Results of NSL dataset –Figure 4 (b)- for all four labels are nearly half of general simulations. U2R false alarm rate in KDD has a fine outcome which is equal to the proposed method. Nevertheless, DR and accuracy of this attack type in general simulations -as pointed up in the previous figure- are not defendable by any means.

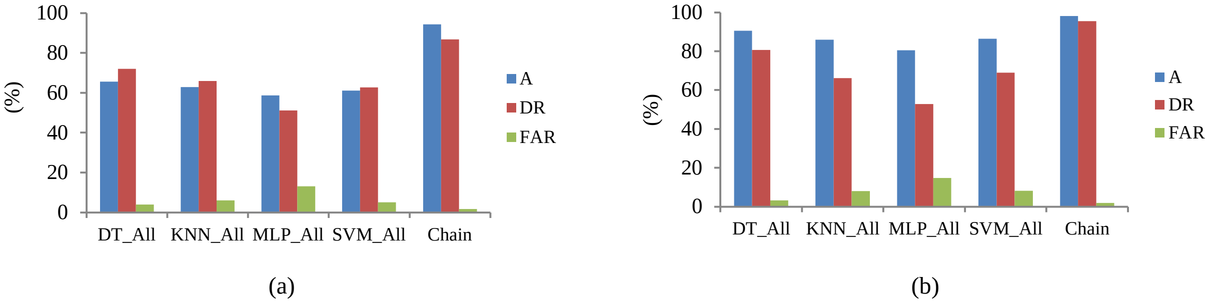

Figure 2. Comparison of average performance of chain method and four classification algorithms

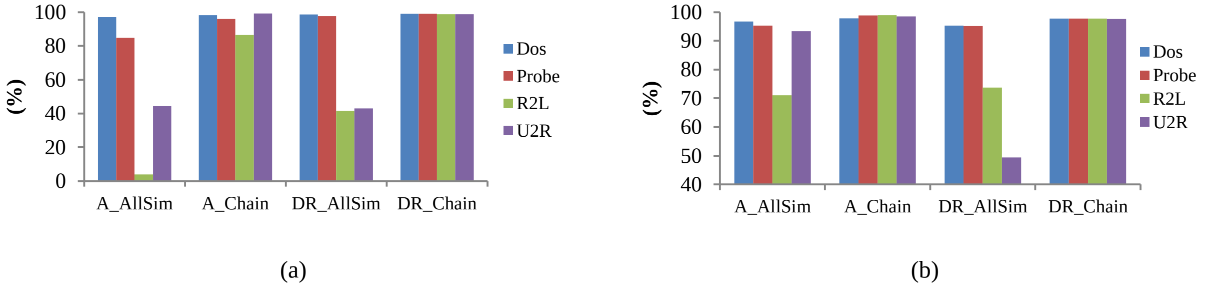

Figure 3. Comparison of best DA and ACC of chain method and general simulations based on different labels

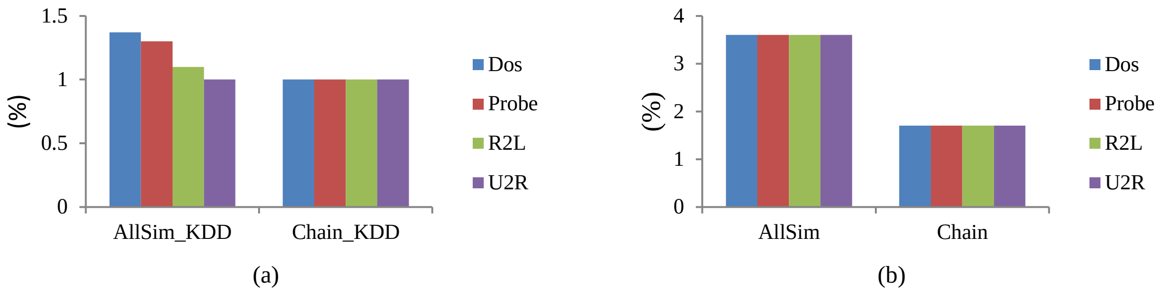

Figure 4. Comparison of best FAR of chain method and four classification algorithms based on different attack types

For better evaluation of proposed method, it is required to compare the chain method to recent studies. For every dataset, DR of every label has been provided in favor of comparison. In paper [26], it has provided the outcomes of simulation in both datasets while Fossaceca et al. [24] worked merely on KDD and Bamakan et al. [25] on NSL. Table 2 shows the performance of proposed chain method versus these studies.

As illustrated in Table 2, proposed method has superior performance in comparison with other studies, especially in R2L and U2R detections. In every label for every dataset, best results have been bolded. Proposed method's Dos and Normal performances are inferior to other methods however their differences are far lower in comparison to R2L and U2R. However, for overall observation of the method performance, the average results of every label could be measured as a more reliable degree for comparison. The average performance of both dataset's labels suggests a considerable enhancement in comparison with recent studies. IDS with high-quality detection ability of all attack types might seem more supportive rather than IDS with outstanding detection ability for some attacks kinds and poor detection capability for other dangerous attack types. Therefore, the real world applications can suggest proposed chain method as a more reliable IDS rather than three compared papers.

5. SEQUENCE DISCUSSION

The sequence of attack types as chain module input is truly essential. The Table 3 shows the proposed sequences for every attack type.

In Table 3, every row is related to a specific label. In every attack type, first row belongs to the KDD and second row belongs to NSL dataset. Every row explains the best sequence of attack types for diagnosing an attack label. For example, for DOS detection in KDD, the sequence of 2,3,1,4 which means at first Probe, then R2L, Dos and U2R, has the best detection rate in chain module. Hence, this sequence will be suggested for DOS detections in application of chain module. For some attack types, there is a consistency between both datasets. For example, for U2R the obtained results suggest Dos and Probe in the end of the chain while R2L and U2R have been suggested for the first or second stage of chain module. As claimed before, the chain method improves the detection rate and accuracy of R2L and U2R more notably.

TABLE 2. Comparison of DR of proposed method with 3 recent studies

Dataset	Methods	Dos	Probe	R2L	U2R	Normal	Average
KDD	[11]	**99.9**	97.4	94.9	62.8	**99.9**	91
	[31]	99.8	96.7	87.7	73.4	98.6	91.2
	CAHIN	98.9	**98.9**	**98.8**	**98.8**	98.8	**98.9**
NSL	[31]	**98.9**	96.9	87.9	72.5	98	90.8
	[10]	98.8	89.2	75	59.6	**99.1**	84.3
	CHAIN	97.7	**97.6**	**97.6**	**97.5**	96.9	**97.5**

TABLE 3. Elite sequences in respect of every label

Labels	Datasets	1st	2nd	3th	4th
Dos	**KDD**	2	3	1	4
	NSL	1	4	2	3
Probe	**KDD**	2	1	3	4
	NSL	4	1	2	3
R2L	**KDD**	2	3	1	4
	NSL	4	3	1	2
U2R	**KDD**	3	4	2	1
	NSL	4	3	2	1
Normal	**KDD**	3	4	1	2
	NSL	3	2	1	4

6. CONCLUSION

This paper has tended to detect attack types based on both KDD CUP 99 and NSL_KDD dataset by means of focusing merely on one attack type at the moment. This paper proposed an innovative method called chain, which focuses on detection of every label in every stage. By working on selected features of both original datasets for every attack types, the degree of concentration for one label detection has been increased. Hence, this study demonstrates there is no need to work on entire features to obtained desired performance. Moreover, this paper indirectly suggests the existence of curse of dimensionality effect in both datasets in the detection of all four attack types. In overall, turning dataset to 2-labed ones in every stage and selecting corresponding GA_ proposed features to the label in every stage, shaped the general idea of the concentration of diagnosing every attack types. The different order of examination of attack types in every stage makes different outcomes. With inspection of 24 possible sequences from four attack types, the elite sequences for both datasets are obtainable. This paper extracted the superior sequences for every attack types as chain module inputs. In this way, based on provided results and information in this paper our proposed chain method is valid and reliable in the real world applications.

7. ACKNOWLEDGMENT

The authors of this work would like to express their gratitude toward UCI machine learning for putting datasets in public access for researchers.

8. REFERENCES

1. Tavallaee, M., Bagheri, E., Lu, W. and Ghorbani, A.A., "A detailed analysis of the kdd cup 99 data set", in Computational Intelligence for Security and Defense Applications. CISDA. IEEE Symposium on, IEEE., (2009), 1-6.
2. Jain, A.K., Duin, R.P.W. and Mao, J., "Statistical pattern recognition: A review", ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, Vol. 22, No. 1, (2000), 4-37.
3. Trunk, G.V., "A problem of dimensionality: A simple example", ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, Vol., No. 3, (1979), 306-307.
4. Erfani, S.M., Rajasegarar, S., Karunasekera, S. and Leckie, C., "High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning", ***Pattern Recognition***, Vol. 58, (2016), 121-134.
5. Wang, W., Guyet, T., Quiniou, R., Cordier, M.-O., Masseglia, F. and Zhang, X., "Autonomic intrusion detection: Adaptively detecting anomalies over unlabeled audit data streams in computer networks", ***Knowledge-Based Systems***, Vol. 70, (2014), 103-117.
6. Kuang, F., Xu, W. and Zhang, S., "A novel hybrid kpca and svm with ga model for intrusion detection", ***Applied Soft Computing***, Vol. 18, (2014), 178-184.
7. de la Hoz, E., Ortiz, A., Ortega, J. and de la Hoz, E., "Network anomaly classification by support vector classifiers ensemble and non-linear projection techniques", in International Conference on Hybrid Artificial Intelligence Systems, Springer., (2013), 103-111.
8. Lakhina, S ,.Joseph, S. and Verma, B., "Feature reduction using principal component analysis for effective anomaly–based intrusion detection on nsl-kdd", (2010).
9. Sindhu, S.S.S., Geetha, S. and Kannan, A., "Decision tree based light weight intrusion detection using a wrapper approach", ***Expert Systems with Applications***, Vol. 39, No. 1, (2012), 129-141.
10. Kim, G., Lee, S. and Kim, S., "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection", ***Expert Systems with Applications***, Vol. 41, No. 4, (2014), 1690-1700.
11. Saied, A., Overill, R.E. and Radzik, T., "Detection of known and unknown ddos attacks using artificial neural networks", ***Neurocomputing***, Vol. 172, (2016), 385-393.
12. Wang, G., Hao, J., Ma, J. and Huang, L., "A new approach to intrusion detection using artificial neural networks and fuzzy clustering", ***Expert Systems with Applications***, Vol. 37, No. 9, (2010), 6225-6232.
13. Meng, W., Li, W. and Kwok, L.-F., "Efm: Enhancing the performance of signature-based network intrusion detection systems using enhanced filter mechanism", ***Computers & Security***, Vol. 43, (2014), 189-204.
14. Lin, W.-C., Ke, S.-W. and Tsai, C.-F., "Cann: An intrusion detection system based on combining cluster centers and nearest neighbors", ***Knowledge-Based Systems***, Vol. 78, (2015), 13-21.
15. Bukhtoyarov, V. and Zhukov, V., "Ensemble-distributed approach in classification problem solution for intrusion detection systems", in International Conference on Intelligent Data Engineering and Automated Learning, Springer., (2014), 255-265.
16. Catania, C.A., Bromberg, F. and Garino, C.G., "An autonomous labeling approach to support vector machines algorithms for network traffic anomaly detection", ***Expert Systems with Applications***, Vol. 39, No. 2, (2012), 1822-1829.
17. Li, W., "Using genetic algorithm for network intrusion detection", ***Proceedings of the United States Department of Energy Cyber Security Group***, Vol. 1, (2004), 1-8.
18. Fidelis, M.V., Lopes, H.S. and Freitas, A.A., "Discovering comprehensible classification rules with a genetic algorithm", in Evolutionary Computation. Proceedings of the 2000 Congress on, IEEE. Vol. 1, (2000), 805-810.
19. De la Hoz, E., De La Hoz, E., Ortiz, A., Ortega, J. and Prieto, B., "Pca filtering and probabilistic som for network intrusion detection", ***Neurocomputing***, Vol. 164, (2015), 71-81.
20. De la Hoz, E., de la Hoz, E., Ortiz, A., Ortega, J. and Martínez-Álvarez, A., "Feature selection by multi-objective optimisation: Application to network anomaly detection by hierarchical self-organising maps", ***Knowledge-Based Systems***, Vol. 71, (2014), 322-338.
21. Shafi, K. and Abbass, H.A., "An adaptive genetic-based signature learning system for intrusion detection", ***Expert Systems with Applications***, Vol. 36, No. 10, (2009), 12036-12043.
22. Tsang, C.-H., Kwong, S. and Wang, H., "Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection", ***Pattern Recognition***, Vol. 40, No. 9, (2007), 2373-2391.

23. Eesa, A.S., Orman, Z. and Brifcani, A.M.A., "A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems", ***Expert Systems with Applications***, Vol. 42, No. 5, (2015), 2670-2679.

24. Fossaceca, J.M., Mazzuchi, T.A. and Sarkani, S., "Mark-elm: Application of a novel multiple kernel learning framework for improving the robustness of network intrusion detection", ***Expert Systems with Applications***, Vol. 42, No. 8, (2015), 4062-4080.

25. Bamakan, S.M.H., Wang, H., Yingjie, T. and Shi, Y., "An effective intrusion detection framework based on mclp/svm optimized by time-varying chaos particle swarm optimization", ***Neurocomputing***, Vol. 199, (2016), 90-102.

26. Bhuyan, M.H., Bhattacharyya, D. and Kalita, J.K., "A multi-step outlier-based anomaly detection approach to network-wide traffic", ***Information Sciences***, Vol. 348, (2016), 243-271.

27. Amiri, F., Yousefi, M.R., Lucas, C., Shakery, A .and Yazdani, N., "Mutual information-based feature selection for intrusion detection systems", ***Journal of Network and Computer Applications***, Vol. 34, No. 4, (2011), 1184-1199.

28. Sangkatsanee, P., Wattanapongsakorn, N. and Charnsripinyo, C., "Practical real-time intrusion detection using machine learning approaches", ***Computer Communications***, Vol. 34, No. 18, (2011), 2227-2235.

29. Pereira, C.R., Nakamura, R.Y., Costa, K.A. and Papa, J.P., "An optimum-path forest framework for intrusion detection in computer networks", ***Engineering Applications of Artificial Intelligence***, Vol. 25, No. 6, (2012), 1226-1234.

9. APPENDIX

Appendix 1. Proposed feature by GA, KDD/NSL

1	1/0	0/1	0/0	0/0	**22**	0/0	1/0	0/1	1/1
2	1/1	1/0	1/1	0/0	**23**	1/1	1/1	1/0	0/0
3	1/1	1/1	0/1	0/1	**24**	1/0	0/0	0/0	0/0
4	1/1	1/1	1/0	0/0	**25**	0/0	1/0	0/1	1/1
5	0/0	0/1	0/0	1/1	**26**	1/0	1/0	1/0	0/0
6	1/1	1/0	0/1	1/1	**27**	1/0	1/1	1/1	0/0
7	0/1	0/1	0/1	0/0	**28**	1/0	0/0	0/1	0/1
8	0/0	0/1	0/1	0/0	**29**	1/0	1/0	1/1	0/0
9	1/1	0/0	0/1	1/0	**30**	0/1	1/1	1/1	1/0
10	0/0	1/0	0/0	1/1	**31**	0/1	1/0	1/1	1/0
11	1/1	1/1	1/1	0/0	**32**	0/0	1/0	1/0	0/0
12	0/0	1/1	1/0	1/0	**33**	0/1	0/0	0/1	0/0
13	1/0	0/0	0/0	0/0	**34**	0/0	0/0	0/0	1/0
14	0/0	1/1	0/0	0/0	**35**	1/0	0/1	0/1	0/0
15	0/0	0/0	1/1	0/0	**36**	1/1	1/1	0/0	0/0
16	1/0	0/1	0/0	0/0	**37**	1/0	0/1	0/1	1/0
17	0/0	0/0	0/1	1/0	**38**	0/0	0/0	0/0	0/0
18	0/1	0/1	1/0	0/0	**39**	1/0	0/0	1/0	0/1
19	1/1	0/0	0/1	0/1	**40**	1/0	0/0	0/1	0/0
20	0/0	1/1	1/1	0/0	**41**	0/0	1/1	0/0	0/0
21	0/1	0/1	0/1	1/1	-	-	-	-	-

Appendix 2. Sequence matrix; specification of every sequence based different orders of attack types

1	1	2	3	4	**13**	3	1	2	4
2	1	2	4	3	**14**	3	1	4	2
3	1	3	2	4	**15**	3	2	1	4
4	1	3	4	2	**16**	3	2	4	1
5	1	4	2	3	**17**	3	4	1	2
6	1	4	3	2	**18**	3	4	2	1
7	2	1	3	4	**19**	4	1	2	3
8	2	1	4	3	**20**	4	1	3	2
9	2	3	1	4	**21**	4	2	1	3
10	2	3	4	1	**22**	4	2	3	1
11	2	4	1	3	**23**	4	3	1	2
12	2	4	3	1	**24**	4	3	2	1

A Novel Intrusion Detection Systems based on Genetic Algorithms-suggested Features by the Means of Different Permutations of Labels' Orders

J. Ghasemi[a], J. Esmaily[b]

[a] *Faculty of Engineering & Technology, University of Mazandaran, Babolsar, Iran*
[b] *Shahid Rajaee Teacher Training University, Tehran, Iran*

PAPER INFO

Paper history:
Received 09 April 2017
Received in revised form 03 May 2017
Accepted 07 July 2017

Keywords:
Intrusion Detection Systems
Feature Selection
Genetic Algorithms

چکیده

سیستم های تشخیص نفوذ با بهره گیری از روش های یادگیری ماشین، قادر به شناسایی رفتار ترافیک های مشکوک به حمله هستند. به علت وجود تعداد نسبتا زیادی ویژگی در دیتاست های استاندارد این حوزه، مثل دیتاست های KDD و NSL، روش های استخراج ویژگی بسیار می‌توانند مفید بعمل آیند. روش های بهینه سازی ترکیبی مثل ژنتیک نیز قادر به یافتن یک راه حل نزدیک به بهینه در مورد ویژگی های مفید استخراجی هستند. این مقاله یک روش خلاقانه به نام "زنجیر" ارائه می‌دهد. تمرکز اصلی ما توجه به یک نوع حمله در هر مرحله است. در شروع کار، ویژگی های استخراجی توسط ژنتیک بدست می‌آیند. بر اساس دنباله ورودی – که از جایگشت ترتیب برچسب ها بدست می‌آید- یک نمونه از ترافیک وارد تابع زنجیر می‌گردد. در هر مرحله یک برچسب – بر اساس ترتیب وارده- شناسایی می‌شود. در صورت موفقیت در مرحله، نمونه به عنوان حمله شناخته شده و در غیر این صورت نمونه سالم تشخیص داده می‌شود. شبیه سازی انجام شده حاکی از برتری روش پیشنهادی نسبت به روش های سنتی است. روش پیشنهادی قادر به تشخیص حملات پیچیده R2L و U2R با دقت بترتیب ۹۸٫۸۳٪ و ۹۸٫۸۸٪ است.

doi*: 10.5829/ije.2017.30.10a.10*