Classifier Algorithm with Attribute Selection and Optimization for Intrusion Detection System.

IOP Conf. Series: Materials Science and Engineering **662** (2019) 022066 doi:10.1088/1757-899X/662/2/022066

Classifier Algorithm with Attribute Selection and Optimization for Intrusion Detection System.

A R Syarif[1], W Gata[2], M Wahyudi[3], S Humaira[3]
[1] STMIK Nusa Mandiri, Jakarta, Indonesia
[2] Universitas Bina Sarana Informatika, Jakarta, Indonesia

E-mail: arief.rma@nusamandiri.ac.id

Abstract. The purpose of this study is to determine how computer network security needed in line with the increasing number of interconnected networks. At present the attacks on computer networks continue to increase, so an efficient network intrusion detection mechanism is needed. Data mining methods over the past few years have been very popular for use in detecting network intrusion. In this paper, we propose reduce the dimensions of the datasets in the preprocessing step by using different state-of-the-art dimension reduction techniques, Principal Component Analysis (PCA) and Information Gain (IG). Particle swarm optimization (PSO) and Genetic Algorithm (GA), both used to find more appropriate set of attributes for classifying intrusion, and *k*-Nearest Neighbors (*k*-NN) algorithm is used as a classifier. The results of the experiments we conducted used the KDD99Cup dataset standard, showing a comparative level of accuracy from the use of dimension reduction and classification optimization. The use of reducing the IG dimension in the KDD99Cup dataset with *k*-NN based PSO optimization can be better at detecting intrusion than other methods.

1. Introduction

In today's modern era so is the rapid development of the Internet, Network security is a major focus in research [1] [2], one of them is about Intrusion Detection System (IDS) as a close relationship in the use of network security services. American computer scientist Rebecca Gurley Bace describes IDS as a computer system (software / hardware) that is used to detect suspicious activity in a network system and conduct analysis to find evidence of intrusion experiments [3].

One of the main causes of network infiltration is the lack of configuration (bugs) in the software used by computers, so that each system requires a method that is tasked with detecting attacks on network security systems [4]. Intrusion Detect on System (IDS) is a method that will be used to identify and track attacks. Anderson James P introduced the first concept of IDS in 1980 [5]. In 1984 Fred Cohen said that increasing attacks on computer networks would increase data traffic [6]. Dorothy E. Denning in 1986 introduced the IDS model that was the basis of the current IDS Model [7]. The higher the level of attack, the right consideration is needed when creating the IDS method. Although there are still problems with IDS, such as low detection of new types of attacks, high-level false alarms and lack of efficiency are obtained when identifying attacks, but IDS is very necessary in an effort to maintain the network system and reduce damage caused by attacks [8]. An attack can be considered a very serious attack if it threatens computer security policies (threatening the confidentiality and integrity of company data).

In the research, a Data Mining is needed which is an alternative in collecting data with much better time and cost efficiency, so as to get a pattern that has mutual interplay in large data sets. KDD99Cup is one of the results of the Data Mining process. In a previous study based anomaly detection such as Lee et al. [9] discussed the selection of features based on the Decision Tree and Genetic Algorithm (GA) combined with the Naïve Bayesian network, using KDD99Cup as an experimental dataset. In addition, several other researchers also used KDD99Cup as an evaluation material for the intrusion approach [10]. Even though using the KDD99Cup as training data can reduce the error rate in minority attacks, but to meet real-time attack detection studies is still not possible. For this study, we will compare algorithms using Principal Component Analysis (PCA) as an attribute that will select and optimize attributes to get the best accuracy before classification. The algorithms used include k-Nearest Neighbors (k-NN), k-NN-based Particle Swarm Optimization (PSO) as decision makers and k-NN-based Genetic (GA) Algorithms as complex optimization problems so that we can know which one has a better level of accuracy and efficiency. Besides learning machine techniques using several algorithms have also been used as a method of detecting anomalies such as Neural Networking [11], Learning Rule [12], Statistics [13] and so on. The method approach using the k-Nearest Neighbors (k-NN) algorithm can be said to be an effective classification of IDS, also used to classify types of attacks and reduce FPR [14], False Alarm Filtering Using k-NN Classifier [15], and use neutral PSO and k-NN binary algorithms without reducing the level of security [16] Studies reveal that modern IDSs find it difficult to handle high-speed network traffic [17]. The researchers also revealed how attackers could exploit this weakness to hide their exploits. They do this by using foreign information to overload the IDS when they carry out attacks [16].

In this paper, we put forward the pre-process data pipeline, using data mining and strategies for reduction features, PCA and IG. The experimental results show that we have compared the efficiency and accuracy of several algorithm methods with reduced features.

2. Method

The hybrid algorithm method used is a comparison between the k-NN, PSO and GA algorithms to get the best set of attributes so that we can get a better level of accuracy and time. At this stage, a little explanation will be given about the k-NN, PSO and GA algorithms, then explain the technique that will be proposed.

2.1 k-Nearest Neighbors (k-NN).

This classification technique with the k-NN algorithm, based on an analogy with the ratio of records of test results given from the same training. Each recorded result is a point in the dimension space n. n represented as an attribute of the results of the training note. Each result of the training note will be stored in the dimensional pattern space, and when an unknown record is detected, the k-NN algorithm will look for a special pattern from the results of the nearest training record. Unrecognized recordings were obtained from the highest achievement of $k's$ own training, k- "Nearest-Neighbors". "Proximity" itself can be interpreted as a measure of distance metrics. Euclidean distance between points or notes if,

$$x_1 = (x_{11}, x_{12} \dots x_{1n}) \tag{1}$$

$$x_2 = (x_{21}, x_{22} \dots x_{2n}) \tag{2}$$

And the approach used,

$$dist(x_1, x_2) = \sqrt{\sum \frac{n}{i=1} (x_1 - x_2)^2} \tag{3}$$

2.2 Particle Swarm Optimization (PSO).

This technique is able to obtain an overall solution optimally in the search space through relationships between individuals in groups of particles. PSO is derivative-free, zero-order method. That means it does not need gradients, so it can be applied to a variety of problems, including those with discontinuous or non-convex and multimodal problems.

The pseudo code of the PSO algorithm is defined as:

1. **procedure** PSO
2. **for** particle $i \in \{1,2,\dots,N\}$ **do**
3. **for** dimension $j \in \{1,2,\dots.,n\}$ **do**
4. **set** $s_{ij} \sim U(lowerBoudary_j, upperBoundary_j)$
5. **set** $d_j \longleftarrow \left|upperBoundary_j - lowerBoundary_j\right|$
6. **set** $v_{ij} \sim U(-d_j, d_j)$
7. **set** $p_{ij} \longleftarrow s_{ij}$
8. **if** $f(p_{ij}) \lessdot f(p_{gi})$ **then**
9. **set** $p_{gi} \longleftarrow p_{ij}$
10. **for** timestep $t \in \{1,2,\dots,I_{max}\}$ **do**
11. **for** particle $i \in \{1,2,\dots,N\}$ **do**
12. **set** $r_p \sim U(0,1)$
13. **set** $r_g \sim U(0,1)$
14. **for** dimension $j \in \{1,2,\dots,n\}$ **do**
15. **update** v_{ij}, s_{ij} from $(3), (4)$
16. **if** $f(s_{ij}) < f(p_{ij})$ **then**
17. **set** $p_{ij} \leftarrow s_{ij}$
18. **if** $f(p_{ij}) < f(p_{gi})$ **then**
19. **set** $p_{gi} \leftarrow p_{ij}$
20. **print** best solution

2.3 Genetic Algorithm (GA).

This algorithm is based on biological process behaviour, which can be used mostly for problem search optimization.

Generic Genetic Algorithm,

1. Generate the initial population collection.
2. While the stop criteria are not reached do.
3. For each chromosome in the population do.
4. Calculate chromosome fitness.
5. Choose chromosomes for crossover.
6. Perform crossover.
7. Make a Mutation.
8. Change populations with new chromosomes.
9. Restore the most suitable chromosome.

2.4 Description of the method used

1. Attribute selection
2. Classification & Optimization.

The research this time uses a proposal by making a one-to-one approach to classify and optimize each attack detection that occurs. Like, when detecting a normal record, it is used as class 1, and for all other types of attacks is set to class 2. Then the attribute is selected using PCA and IG, while to get the classification class uses the k-NN algorithm and for optimization PSO and GA. Overall the contents of the composition can be seen in Figure 1.

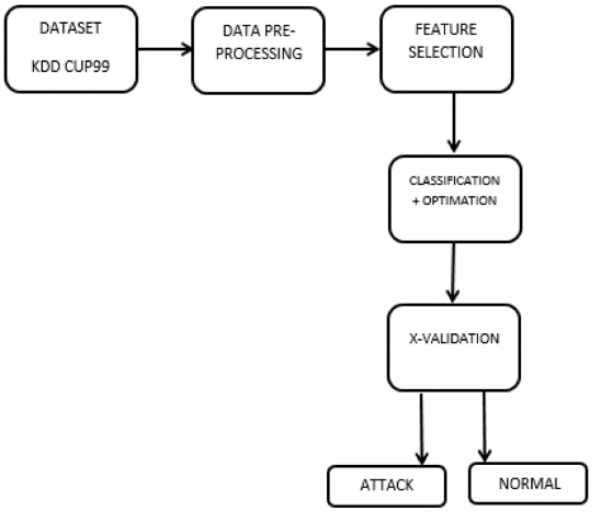

Figure 1. Method used.

2.4.1 Dataset.

Data Collection of Data Discovery and Mining Knowledge KDD99Cup is general data that has been widely used by researchers to make a simulation in the Intrusion Detection System (IDS) [18] so that when the process of measuring technical performance can be more focused and objective IDS core. This study uses a KDD99Cup dataset, which has 9.373 records, 30 attributes (features) and 1 attribute as the last label. There are 30 attributes that are used and consist of 9 basic types, 21 types of content and the rest of the types of traffic in a network.

2.4.2 Preprocessing Data.

2.4.2.1 Attack Category.

At the data processing stage, there will be 5 categories of 23 numbers of types of attacks contained in the KDD CUP99 dataset. The types of categories contained in the KDD CUP99 dataset include: Normal, DoS, PROBE, U2R and R2L like the information shown in the Table 1 below:

Table 1. Category in KDD99Cup dataset

Category	Type of Attack
Normal	Normal Connection
PROBE	smurf , teardrop, Back, Neptune, pod, Land
U2R	Rootkit, perl, Buffer_overflow, loadmodule
R2L	imap, multiloop, warezclient, warezmaster, guess_passwd, phf, ftp_write, spy

PROBE, is an attack that scans the network to find security holes. An attacker who knows the sensitivity of a server connected to the network can exploit the server. This type of attack is widely applied to social engineering techniques.Remote-to-Local (R2L), is an attack carried out by an attacker that passes through the network to the server to find security holes and creates an account to enter. Denial-of-Service (DoS) is an attack that causes the server to be busy due to a processing request so it cannot provide the services needed by the actual user. Like, when a user requests a Hypertext Transfer Protocol (HTTP) service from a server through a web browser, a service denial occurs so that the service is not accessible. User-to-Root (U2R), the original user uses the system on

the server to get access because in this type of attack the administrator is the type of attack. The method that is mostly used in User-to-Root attacks is Buffer overflow.

2.4.2.2 Sample

The next step in the data processing stage is to use a random sample of 9.373 notes contained in the KDD99Cup dataset by noting that all types of attacks are in the sample record. The purpose of data reduction or data reduction by taking notes of random samples from the dataset of KDD99Cup is to improve manufacturing and reduce training time, size and complexity of the model (See Table 2).

Table 2. Attack Category

Attack category	Number of records
Normal	3000
DoS	2764
PROBE	2431
U2R	52
R2L	1126
Total records	9373

2.4.3 Attribute Selection

2.4.3.1 Principal Component Analysis (PCA)

PCA is an attribute selection technique that has the task of reducing the number of attributes or features contained in the KDD99Cup dataset before entering into the classification process using the k-Nearest Neighbors (k-NN) algorithm based on Particle Swarm Optimization (PSO) and Genetic Algorithm (GA). This technique is used to get as many as 31 attributes from 41 attributes in the KDD99Cup dataset for each type of attack category.

2.4.3.2 Information Gain (IG)

IG evaluates attributes by measuring their information acquisition with respect to class. Discretizing numeric attributes first uses MDL-based discretization methods.

3. Results and Discussion

In this section, we show our experimental results for detecting intrusion using attribute reductions and classification algorithms using the KDD99Cup dataset by creating training and testing sets. All experiments were carried out as 10 times with 10-fold cross-validation. TP, FP, TN and FN are variables used to measure the accuracy of intrusion detection system, where TP states that normal behaviour is predicted correctly, FP indicates that abnormal behaviour is assessed as normal, FN indicates that wrong behaviour is usually considered abnormal, and TN shows abnormal behaviour detected correctly.

Accuracy is the ratio of correctly classified to the total classified examples.

$$Acc = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (4)$$

Where, FN is False Negative, TN is True Negative, TP is True Positive and FP is False Positive.

The research in this paper uses Rapidminer 9.0 software installed on notebook hardware with technical specifications, AMD A10-9620 CPU 2.50 GHz and 8.0 GB RAM. 4.2. Since it is not possible to carry out experiments with whole KDD99Cup dataset because of dimensionality, existence of redundancy and class imbalance, therefore, a compact dataset was formed manually in which all 41 features are present. In second step, compact dataset was imported and attribute ranking and selection algorithms were applied. Information gain (IG) algorithm was applied by calculating entropy of each attribute to know extent of information present in different features of dataset. The process of reducing

an attribute in the attribute selection technique can use the Principal Component analysis (PCA) technique (see Table 3).

Table 3. Comparison Results.

	IG	IG + GA	IG + PSO	PCA	PCA + GA	PCA + PSO
Normal	99,36%	99,91%	99,91%	99,37%	99,91%	99,73%
DoS	99,19%	99,87%	99,77%	99,19%	99,88%	99,88%
PROBE	99,51%	99,64%	99,72%	99,51%	99,62%	99,64%
U2R	99,62%	99,74%	99,75%	99,62%	99,72%	99,74%
R2L	99,14%	99,17%	99,74%	99,14%	99,29%	99,26%

Seen from the table above, we have done 6 times experimental to calculate the accuracy of the classification results of detecting network intrusion for each category. The first experimental method shows that accuracy with attribute selection, IG and classification using k-NN the algorithms from the KDD99Cup dataset show that the method can detect the intrusion of the U2R category better than the other categories. In the second experimental, DoS category intrusion has a better level of accuracy than other categories using IG selection attribute and k-NN algorithms that have been optimized with GA. The interesting thing from the third experimental is that the PROBE, R2L, U2R and Normal categories have almost the same accuracy even though the DoS category still has a better level of accuracy using the IG selection features and k-NN algorithms that have been optimized with PSO.

All four to six experimental use PCA features selection with k-NN classifiers algorithms that have been optimized with GA and PSO. The normal category has a very low level of accuracy when the fourth experimental is compared to previous experimental. The PROBE category has the accuracy of the classification results that are better than the other categories in the fifth and sixth experimental. (See Figure 2).

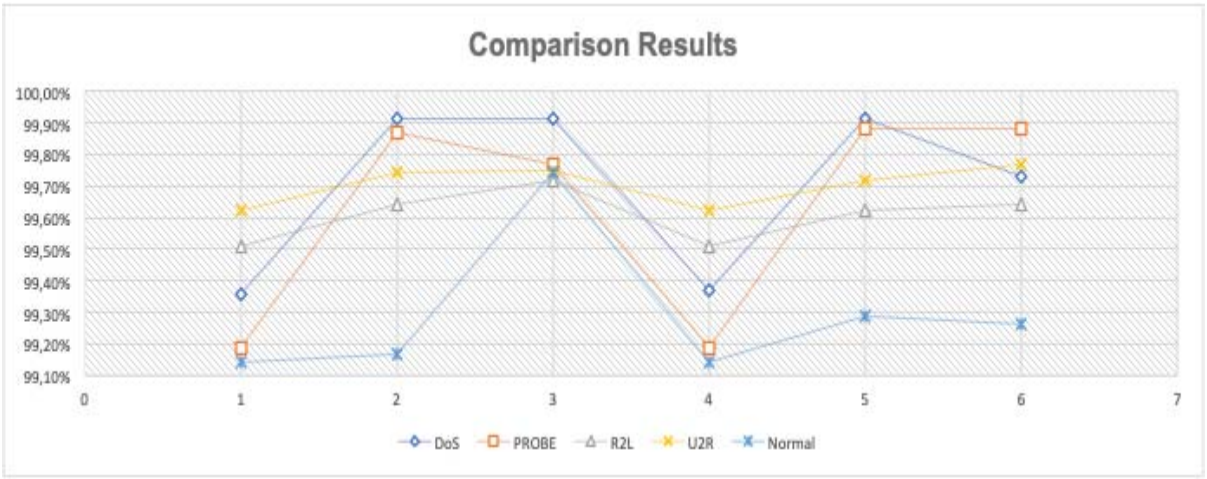

Figure 2. Comparison results.

As shown in Figure 2, from the results of the experimental that have been carried out it can show that the features extracted by IG can provide more additional discriminatory information to improve the accuracy of the classification algorithm compared to PCA. Reducing the dimensions of the KDD99Cup dataset can also improve the performance of generalizations and running time of the k-NN classifier. The overall results of the experimental show that IG is better than PCA.

Table 4. IG-k-NN-PCA for DoS Category

DoS	Attibute selection	
	IG	PCA
k-NN	99,36%	99,37%
k-NN GA	99,91%	99,91%
k-NN PSO	99,91%	99,73%

We can also see from Table 4 that the stabilities of the learning of IG-k-NN-GA, IG-k-NN-PSO, PCA-k-NN-GA and PCA-k-NN-PSO are better than IG-k-NN and PCA-k-NN. In addition, this table can also see that the overall performance of IG-k-NN-GA model is better than other methods for intrusion detection.

4. Conclusion

We can conclude that the results of accuracy and efficiency in time can use the attribute selection IG method with $\boldsymbol{k}$-NN based PSO algorithm can get better results detect intrusion degree of accuracy than using other methods. This conclusion can be obtained see the results of the research that was made before in the R2L category using PCA and classifying $\boldsymbol{k}$-NN based PSO with a dataset of 9.373 records and 30 attributes, where $\boldsymbol{k = 5}$ can only get a percentage of 99,26%, compared to weights of the attribute use IG and classification of $\boldsymbol{k}$-NN based PSO get a better accuracy rate of 99,74%. So, the results of the experimental we have done, that by using attribute selection, IG and $\boldsymbol{k}$-NN classification based PSO, we can get an accuracy rate of 0,48% and time efficiency that is far better than other methods. Although in some attack categories there are differences in the level of accuracy, as in the U2R attack category experiment who used the PCA selection attribute and $\boldsymbol{k}$-NN based PSO classification had a difference of 0,02% with the DoS attack category who used future selection, PCA and the $\boldsymbol{k}$-NN classification had a difference of 0,01%, but from the overall experimental results we had done, the attribute selection method using IG can increase the accuracy of the $\boldsymbol{k}$-NN classification based PSO at compare the other methods.

Acknowledgement

The authors would like to thank all member Yayasan Komunitas Open Source who have become participants in this research. The authors also would like to thank our family who provided the endless love and tireless support during the process.

References

[1] Dasgupta, D., Yu, S., & Nino, F. (2011). Recent advances in artificial immune systems: models and applications. Applied Soft Computing, 11(2), 1574-1587.
[2] Aickelin, U., Greensmith, J., & Twycross, J. (2004, September). Immune system approaches to intrusion detection–a review. In *International Conference on Artificial Immune Systems* (pp. 316-329). Springer, Berlin, Heidelberg.
[3] Bace, R., & Mell, P. (2001). *NIST special publication on intrusion detection systems*. BOOZ-ALLEN AND HAMILTON INC MCLEAN VA.
[4] Syarif A R and Gata W 2017 *Intrusion Detection System Using Hybrid Binary PSO and K-Nearest Neighborhood Algorithm*, **181–6**
[5] Anderson J P 1980 *Computer Security Threat Monitoring And Surveillance*
[6] Cohen F 1987 *Computer viruses. Computers and Security*, 6(1), **22– 35**
[7] Denning D E 2012 *An intrusion-detection model*. Proceedings - IEEE Symposium on Security and Privacy, (2), **118–131**
[8] Kuang F, Xu, W and Zhang S 2014 *A novel hybrid KPCA and SVM with GA model for intrusion detection*. Applied Soft Computing Journal, 18, **178–184**

[9] Lee W, Fan W, Miller M, Stolfo SJ, Zadok E. Toward cost-sensitive modeling for intrusion detection and response. J Compute Secure 2002;**10:5–22**.
[10] Su M Y 2011 *Using clustering to improve the KNN-based classifiers for online anomaly network traffic identification.* Journal of Network and Computer Applications, 34(2), **722–30**
[11] Cho S, and Park H 2003 *Efficient anomaly detection by modeling privilege flows using hidden Markov model*. Computers & Security, 22(1), **45–55**
[12] Lazarevic A, Ertoz L, Kumar V, Ozgur A and Srivastava J. (n.d.) *A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection* † 2 Evaluation of Intrusion Detection Systems, **25–36**
[13] Anderson D, Lunt T F, Javitz H, Tamaru a and Valdes a 1995 *Detecting unusual program behavior using the stastistical component of the next-generation intrusion detection expert system (NIDES)*. Computer Science Laboratory SRI-CSL, (910097**), 6–95**
[14] Liao Y and Vemuri V R 2002 *Use of k-nearest Neighborsclassifier for intrusion detection. Computers and Security,* 21(5), **439–448**
[15] Law K 2005 *IDS false alarm filtering using KNN classifier*. 5th International Workshop WISA, Revised Selected Papers, **114–121**
[16] Mukherjee S and Sharma N 2012 *Intrusion Detection using Naive Bayes Classifier with AttributeReduction*. Procedia Technology, 4, **119–128**
[17] V Paxson 1999 "*Bro: a system for detecting network intruders in real-time*," Comput. Networks, vol. 31, no. 23–24, pp. **2435–63**
[18] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "*A detailed analysis of the KDD CUP 99 data set,*" IEEE Symp. Comput. Intell. Secur. Def. Appl. CISDA 2009, no. Cisda, pp. **1–6**