IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 24, No. 5, October, 2019, pp. 513-518
Journal homepage: http://iieta.org/journals/isi

Skyline Computation for Improving Naïve Bayesian Classifier in Intrusion Detection System

Abdelkader Alem[1*], Youcef Dahmani[2], Bendaoud Mebarek[3]

[1] Ecole Superieure Nationale d'Informatique, ESI, BP 68 M 16270 Oued Smar Algiers, Algeria
[2] University of Tiaret, EECE Lab, BP 78 Zaaroura Tiaret, 14000, Algeria
[3] The Research Laboratory of Industrial Technologies, University of Tiaret, 14000, Algeria

Corresponding Author Email: a_alem@esi.dz

https://doi.org/10.18280/isi.240508

Received: 15 May 2019
Accepted: 20 August 2019

Keywords:
network security, intrusion detection system, naïve Bayesian network, skyline operator

ABSTRACT

Intrusion detection systems (IDSs) are critical to network security. However, there are some common defects with the existing IDSs, namely, low detection rate of rare attacks and high number of false alarms. Many have suggested solving these defects by integrating different IDSs techniques, but the effectiveness has not been justified. This paper puts forward a two-layer hybrid IDS based on Skyline operator and Naïve Bayesian classifier. First, the most suitable classifier was identified through Skyline computation based on three criteria, namely, accuracy, detection rate and false alarm rate. Then, the results were integrated by the Naïve Bayesian classifier into the final decision. To verify its effectiveness, the proposed IDS was tested on the famous KDD dataset. The results show that our system greatly improves the detection rate of rare attack, while decreasing false alarms rate, from the levels of the previous techniques.

1. INTRODUCTION

Nowadays, for a few dollars, hacking tools and computer attacks are provided by experts to amateurs. These tools vary according to their dangerousness and performance and some of them could bypass the most sophisticated security mechanisms. A successful attack may cause very serious losses depending on purpose, magnitude, and dangerousness. Faced this, security has become an obligation that allows the protection of information and information systems against any threat. Moreover, despite all security mechanisms that could be put in place, certain types of attacks could bypass them. In order to enhance the security of computer systems and network, intrusion detection system (IDS) concept was introduced [1, 2].

An IDS is a tool that allows us to predict or identify any unauthorized activity in a network. However, IDSs also have some weaknesses. Indeed, some attacks may go unnoticed (false negatives), or some alerts may be generated against attacks that have not occurred (false positive). In addition, the number of alerts generated is often too high so that the security operator who is responsible for analyzing and processing these alerts is quickly drowned. In this context we distinguish two approaches to detect intrusions: anomaly detection and misuse detection (signature detection). The first consists in searching known signatures of attacks while the second consists in defining normal behaviour of the system and determining a baseline separator between normal and abnormal behaviour. This could infer any deviation from the normal profile as hypothetical attack. Several hybrid data mining techniques were developed for improving IDS performances; such as Naïve Bayes [3, 4], Support Vector Machines [5, 6], and decision trees [7]. The justification for this hybridization is not always given. In the present work, a proposal is made based on Skyline's computation which will select the best classifiers to be combined. The remainder of this paper is outlined as follows: section 2 presents a background and related work; Section 3 presents the proposed model. Next, the experimental results are discussed in Section 4. Finally, section 5 concludes the paper.

2. BACKGROUND AND RELATED WORK

2.1 Skyline queries

Skyline computation is applicable in many applications that require multi criteria decision making. The term skyline operator was introduced by Borsony et al. [8], the computation of Skyline queries is also known as the Pareto optimum or the maximum vector problem [9].

According to Pareto dominance relationship, Skyline queries select all promising (non-dominated) instances from multi-dimensional dataset. Let D a set of d-dimensional points, a Skyline statement, retrieves, the Skyline, set of points representing the good combination (not dominated) of all criteria. A point p dominates another point q, according to Pareto optimality, if and only if p is at least as good as q in all dimensions and strictly better than q in at least one dimension. The skyline points are incomparable. We said p dominates q and we formally write:

$$p \succ q \Leftrightarrow \begin{cases} \forall\, s \in S: s(p) \geq s(q) & and \\ \exists\, s \in S: s(p) > s(q) \end{cases} \quad (1)$$

where, s is a score function.

The Skyline points in D will be the points satisfying:

$$p \in D \mid \nexists p' \in D: p' \succ p \quad (2)$$

2.2 Bayesian networks

The representation of uncertain knowledge is an important problem in the field of artificial intelligence. Bayesian networks offer an interesting solution for many theoretical and practical issues, they are a formalism of probabilistic reasoning pioneered by Pearl et al. [10], have figure out very user friendly tools for representing uncertain knowledge, and allow reasoning from incomplete information.

Bayesian networks are the combination of probabilistic approaches and graph theory, formally Bayesian network $B=(G,\Theta)$ is depicted as:

- Let a set of observable random variables $X = X_1; :::;X_n$, $G = (X;E)$ a directed acyclic graph (DAG) where each node match a variable of X.
- $\theta = \theta_i = P(X_i/Pa(X_i)$ set of probability distributions of each node Xi given the probability of its parents.

Thus, Bayesian network graph makes a representation in a visual way of the relationship (dependencies and independences) between the variables of the system. The probabilities in a Bayesian network allow representing the uncertain aspect that links the variables. Naïve Bayesian networks are the simplest form of a Bayesian network. The root represents the unobserved node and the leaves the different observations (observed variables) [11].

Naïve Bayes approach is a graph with a single parent, not observed, and several leaf nodes representing observed variables, with a strong assumption of independence between the sheets given their parent. Thus, with a set of training, the only investigation to be performed is the calculation of the conditional probabilities since the network is unique. Once the Bayesian network is quantized, it is possible to classify every new object, given the attribute values utilizing the Bayes rule formulated by:

$$P(C\ /A) = \frac{P(A\ /C).P(C)}{P(A)} \quad (3)$$

2.3 Related work

Horng et al. [12] developed an IDS that integrates a hierarchical clustering algorithm as well as an SVM technique. NFPHIDS [13] is hierarchical IDS constituted of two levels. The first level includes four classifiers: Random Forest, Simple Cart, Best first decision tree and naïve Bayes. Their well known good performance, justify their utilization as ingress data; the second level uses output of the first one that contains Naïve Bayes as final classifier. Ada-Boost algorithm was developed using both Naïve Bayes and decision tree as weak classifiers [14, 15]. Multi-level based IDS composed of the intersection of two different classifiers, fuzzy unordered rule induction algorithm [16] and random forests [17] was proposed by Ahmim et al. [15]. XM-RF [18] is a hybrid IDS based on X-Means clustering and Random Forest classification. First of all, analogous data instances are clustered using X-Means clustering based.

Next, Random Forest classifier is used for rearranging the misclassified clustered data into a new cluster. HFIDS [19] is an IDS using fuzzy logic and applied to wireless local area networks. In the latter, a misuse detection module is connected to the anomaly detection module and the overall decision is performed by fuzzy rules. An IDS based belief function was proposed, it is composed of three stages [20]. At the first one, two detection modules (SVM and Naïve Bayes classifier) have been used. The outputs of the first stage are fuzzified in the second stage. The last stage uses belief function to perform the final decision of the system. With this approach, the result of false alerts is high because it did not take into account conflictual cases between classifiers. Output that is slightly abnormal is considered abnormal, which is not always true.

Ahmim et al. [21] propose an IDS build on probability prediction combination obtained from a tree of classifiers. This IDS contains two layers, the first one is a binary tree of classifiers; in the second layer, a combination of predictions of the first layer is done. This work has the disadvantage of the strong dependence of the choice and order of the classifiers. Any change in this order could lead to different results and therefore making different decisions.

In most previous works, no justification for the choice of classifiers is argued. The present work focuses on the Skyline operator to choose the set of the best classifiers to be combined. Indeed, this work aims to build a high performance and effective intrusion detection system by cooperating and integrating several modules of detection (classifiers) in a naïve Bayesian network to minimize the false alarm rate FAR and increase the detection rate DR of infrequent attacks keeping their high efficiency on the other attacks and the normal behavior.

Skyline computation will be used to choose the best classifiers according to three main criteria which are Accuracy, DR and FAR. Therefore, Skyline classifiers will be considered as a node in the naïve Bayesian network. This network is known by its linear complexity which justifies its choice in this work.

3. NAÏVE BAYES IDS BASED ON SKYLINE

3.1 Overview

The aim of our work is to improve the detection rate on rare attacks and build an efficient Hybrid IDS based on a Skyline of classifiers that will provide good performances. There are different methods and techniques to distinguish the various types of classifiers in data mining. Each classifier could rank every network connection as either a normal behavior or an attack with different error rate. The performance of the different type of classifiers is measured by its ability to classify each connection in the right category. Table 1, known as the confusion matrix, shows the four possible cases:

- True positive (TP): an attack data identified as an attack;
- True negative (TN): a normal data identified as normal;
- False positive (FP): a normal data identified as an attack;
- False negative (FN): an attack data identified as normal.

The commonly used metrics of performance of an IDS are based on three factors (accuracy, detection rate and false alarms rate):

Table 1. Confusion matrix

		Predicted Class	
		Normal	Abnormal
Actual Class	Normal	True Negative	False Positive
	Abnormal	False Negative	True Negative

- $Accuracy = (TP + TN)/(TP + TN + FP + FN)$

- $Detection\ Rate\ DR = (TP)/(TP + FN)$
- $False\ Alarm\ Rate\ FAR = (FP)/(FP + TN)$

3.2 Main structure of the model

The main structure of the proposed model is composed of two levels: the first one includes the best classifiers that are not worse by any other classifier based on the three dimensions (accuracy, detection rate and false alarm rate). The second level contains Naïve Bayesian classifier that integrates the outputs of the first level to make the ultimate decision. The choice of the best classifiers in the first level was based on a Skyline computation (Skyline operator) on three main criteria:

- Maximize the accuracy;
- Maximize the detection rate;
- Minimize the false alerts rate.

As illustrated by the Figure 1, the proposed approach consists in integrating best classifiers predictions as another necessary observation in a naïve Bayesian network to make a good decision.

In the naïve Bayes of Figure 1, the variable Class has two instances (Normal and Abnormal) and leaf nodes .A_1;A_2; ::and A_{39}/ represents the values of attribute connection while the variables C_1; C_2 represents the decision of classifier C_1; C_2 about the variables .A_1;A_2; ::and A_{39}/ and can take the values Normal or Abnormal. In this case classification consists in determining the most probable instance of the class variable Class for an instance of the observed vector of attributes A_1; A_2:::A_{39}; C_1; C_2 by applying the rule presented in section Bayesian Network.

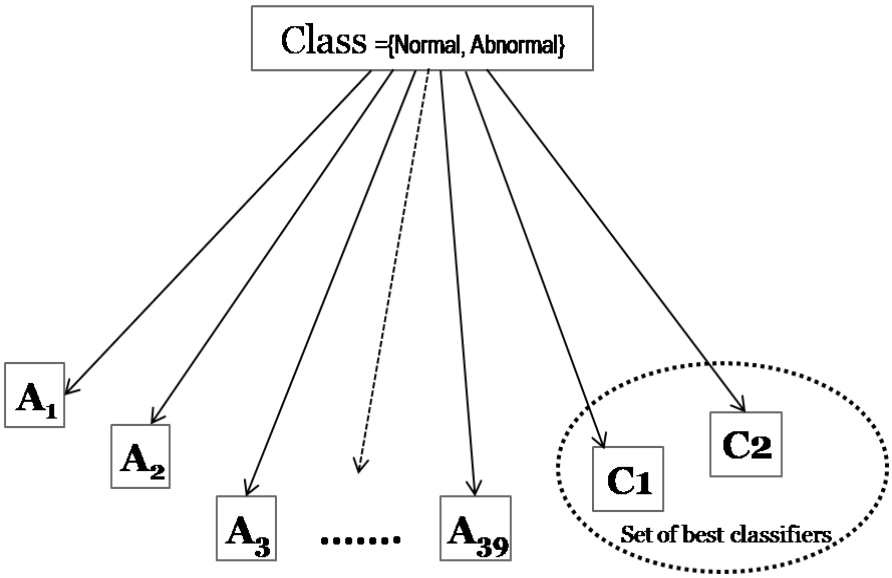

Figure 1. Naïve Bayesian network based on a skyline of classifiers

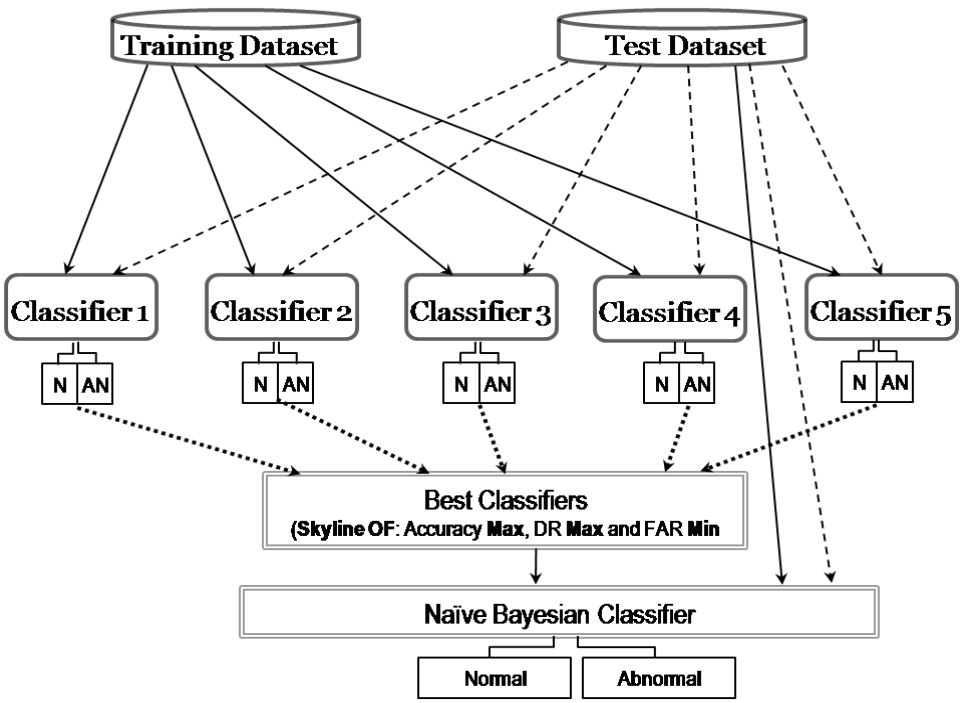

Figure 2. General structure of the proposed model

3.3 Building method

The main idea for the construction of our model is the use of several data-mining techniques and Naïve Bayesian network that combining classifiers outputs and initial observed variables (attributes describing a connection) to determines the most probable class (Normal or Abnormal).

As illustrated in Figure 2, our model is composed of two levels: In the first level, we carried out a comparative study of several data mining techniques in order to select the best classifiers. According to Pareto dominance relationship, we choose the best classifiers that have the better DR, The better of accuracy and with a low FAR as possible.

A comparative study was shown in Table 2 between six data mining techniques: Simple Cart SC, Decision Tree DT, Naïve Bayes NB, Random Forest RF, Best First Tree BFT and Random Tree RT.

Table 2. Comparative results of classifiers

Classifier	DR	Accuracy	FAR
NB	89.21%	92.68%	5.18%
SC	**98.80%**	94.87%	7.53%
RF	94.28%	**95.29%**	**4.08%**
BFT	98.70%	94.87%	7.53%
DT	93.26%	94.65%	4.50%
RT	93.70%	94.78%	4.55%

As is presented in Table 2, SC dominates (in Pareto sense) BFT and RF dominates NB, DT and RT. We get the Simple Cart and Random Forest as incomparable because SC is better than RF with respect to the DR while the RF is better than SC with respect to the accuracy and FAR. So the Skyline of classifiers is SC; RF. After choosing the best classifiers, we add the classifiers outputs as other attributes in the initial training dataset as illustrated in Table 3, for each connection of the initial dataset, we give the decision Normal or Abnormal that represent the outputs of the selected classifiers.

Table 3. Total attributes of the Naïve Bayes IDS

	Initial attributes				New attributes	
	A1	A2	...	A39	**RF**	**SC**
R1	0	tcp	...	1	Normal	Abnormal
R2	0	udp	...	0	**Normal**	**Normal**
R3	1	tcp	...	0.5	Abnormal	Abnormal
...	...	...	...	...	...	...

Table 3 shows that record R1 is represented by all initial attributes and two decisions of RF and SC according the initial attributes.

4. EXPERIMENTS AND RESULTS

The present section consists of two parts. The first one describes the data set used in our experiments. The second represents a comparative study between the proposed model and some related recent works.

4.1 Data set description

In this subsection, we present the dataset used to analyse and evaluate the performance of Skyline-based IDS (S * IDS), we use the well known KDD'99 dataset [22] that represent the most used data for IDS.

The KDD'99 data set was used to analyze and evaluate our method. KDD'99 was originated in 1999 from DARPA-Lincoln98 data set by MIT's Lincoln laboratory. It is organized into five categories: DOS attack, U2R attack, Probe attack, R2L attack and Normal behavior. Every single record of KDD'99 data set has 41 attributes (34 numeric and 7 symbolic). According to [18], the KDD'99 training data set contains normal behaviors and 22 attacks with 4,940,000 data records. The test data set contains 311,029 data records, covering normal behaviors and 37 attacks. It should be noted that 17 of the test data set attacks are not in the training data set. KDD'99_10% is 10% of KDD'99 training data set with the same distribution of attacks and normal behaviors. Table 4 shows the distribution of the attacks and normal behaviors in both KDD'99training 10% and KDD'99 test.

Table 4. KDD'99_10% Data set description

Number of records	Training data set		Test data set	
Category of connection	All	Distinct	All	Distinct
Normal	97278	87832	60593	47913
DOS	391458	54572	229853	23568
Probe	4107	2130	4166	2678
R2L	1126	999	16189	2913
U2R	52	52	228	215
All	494021	145585	311029	77287

We reduced the size of KDD'99_10% by removing all redundant records, thus creating our own training data set. Evaluation of the performance of our model is performed on KDD'99 test data set [22].

4.2 Results and discussion

In this work, WEKA Data Mining Tools [23] has been used to implement both classifiers. The results were obtained using a PC operating on Microsoft Windows OS and equipped with a Processor Intel R Core TMi5 2,4 GHz CPU and 4 GB RAM. To evaluate the performance of our approach as well as to improve the Naïve Bayesian classifier using Skyline computation in intrusion detection utility, we have compared the obtained results with those mentioned by Ahmim and Ghoualmi-Zine [13]. All these works referenced in Table 5 have used KDD'99 to evaluate the performance of their solutions. In this study, our model is built and trained by the parameters mentioned in the section above. Then, all KDD'99 test data set were used as a test data set. As illustrated in Table 5, our approach gives the best detection rate of R2L and U2R without losing a good detection rate compared to works cited in this paper using KDD'99 data set.

As illustrated in Table 5, our approach gives the best accuracy and the second DR without losing a good false alarm rate compared to works cited in this paper using KDD'99 data set.

Note also that our approach has shown its great ability to detect rare attacks such as R2L and U2R. From Table 5, it could be clearly seen that the results obtained by our approach are better than those obtained by other referenced works cited in this paper.

Table 5. Results and performance

	TNR	DR					Accuracy	FAR
		DOS	Probe	R2L	U2R	All		
SC	92.50	92.90	83.90	47.50	13.50	88.90	94.88	7.56
RF	95.90	94.90	63.90	37.80	19.10	94.30	95.29	4.08
NFPHIDS	98.65	97.85	98.13	43.15	72.81	94.26	95.12	1.35
Horng et al	99.30	99.60	97.50	28.80	19.70	94.82	95.70	0.70
HCPTC-IDS	98.87	99.83	95.27	36.50	81.14	95.65	96.27	1.13
S-IDS (our approach)	98.75	97.80	94.70	**93.50**	**96.00**	95.52	**96.97**	1.25

5. CONCLUSIONS

In the present work, a new effective intrusion detection system based on the Skyline computation is presented. The decision-making system is based on combining naïve Bayesian classifier and other data mining techniques. This methodology of selecting the classifiers to be combined was based on the Skyline operator that consists in choosing best classifiers that are not worse by any other classifier on the three dimensions (accuracy, detection rate and false alarm rate). This hybrid technique is suitable in indecision cases and more accurate. The accuracy is due to the fact that incorrect interpretation is very unlikely thanks to the multiple-checking technique in the selected Skyline classifiers. The advantage of the Skyline computation is their ability to select non dominated instances (best classifiers). The output results show that our proposal results in a better performance.

While it gives the best detection rate for rare attacks R2L and U2R, it also preserves a high detection rate and accuracy. This is true even when compared to well-known works in the literature using exactly the same training and test dataset.

REFERENCES

[1] Evangelista, T. (2004). Les IDS: les systèmes de détection d'intrusions informatiques. Dunod.

[2] Wu, S.X., Banzhaf, W. (2010). The use of computational intelligence in intrusion detection systems. A review, Applied Soft Computing, 10(1): 1-35. https://doi.org/10.1016/j.asoc.2009.06.019

[3] Scott, S.L. (2004). A Bayesian paradigm for designing intrusion detection systems. Computational Statistics & data Analysis, 45(1): 69-83. https://doi.org/10.1016/S0167-9473(03)00177-4

[4] Ben-Amor, N., Benferhat, S., Elouedi, Z. (2004). Naive Bayes vs decision trees in intrusion detection systems. Proceedings of the 2004 ACM Symposium on Applied Computing, pp. 420-424. https://doi.org/10.1145/967900.967989

[5] Mukkamala, S., Janoski, G., Sung, A. (2002). Intrusion detection using neural networks and support vector machines. Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290), Honolulu, HI, USA, USA, pp. 1702-1707. https://doi.org/10.1109/IJCNN.2002.1007774

[6] Zhang, Z., Shen, H. (2005). Application of online-training SVMs for real-time intrusion detection with different considerations. Computer Communications, 28(12): 1428-1442. https://doi.org/10.1016/j.comcom.2005.01.014

[7] Paek, S.H., Oh, Y.K., Lee, D.H. (2006). sIDMG: Small-size intrusion detection model generation of complimenting decision tree classification algorithm. International Workshop on Information Security Applications, pp. 83-99. https://doi.org/10.1007/978-3-540-71093-6_7

[8] Borzsony, S., Kossmann, D., Stocker, K. (2001). The skyline operator. Proceedings 17th International Conference on Data Engineering, Heidelberg, Germany, Germany, pp. 421-430. https://doi.org/10.1109/ICDE.2001.914855

[9] Kung, H.T., Luccio, F., Preparata, F.P. (1975). On finding the maxima of a set of vectors. Journal of the ACM (JACM), 22(4): 469-476. https://doi.org/10.1145/321906.321910

[10] Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Reasoning. Morgan Kaufmann, Burlington.

[11] Kenaza, T., Tabia, K., Benferhat, S. (2010). On the use of naive Bayesian classifiers for detecting elementary and coordinated attacks. Fundamenta Informaticae, 105(4): 435-466. https://doi.org/10.3233/FI-2010-373

[12] Horng, S.J., Su, M.Y., Chen, Y.H., Kao, T.W., Chen, R.J., Lai, J.L., Perkasa, C. (2011). A novel intrusion detection system based on hierarchical clustering and support vector machines. Expert systems with Applications, 38(1): 306-313. https://doi.org/10.1016/j.eswa.2010.06.066

[13] Ahmim, A., Ghoualmi-Zine, N. (2013). A new fast and high performance intrusion detection system. International Journal of Security and Its Applications, 7(5): 67-80. https://doi.org/10.14257/ijsia.2013.7.5.06

[14] Natesan, P., Balasubramanie, P., Gowrison, G. (2012). Improving attack detection rate in network intrusion detection using Adaboost algorithm with multiple weak classifiers. Journal of Information & Computational Science, 9(8): 2239-2251. https://doi.org/10.3844/jcssp.2012.1041.1048

[15] Ahmim, A., Ghoualmi-Zine, N. (2014). A new adaptive intrusion detection system based on the intersection of two different classifiers. International Journal of Security and Networks, 9(3): 125-132. https://doi.org/10.1504/IJSN.2014.065710

[16] Huhn, J., Hullermeier, E. (2009). FURIA: An algorithm for unordered fuzzy rule induction. Data Mining and Knowledge Discovery, 19(3): 293-319. https://doi.org/10.1007/s10618-009-0131-8

[17] Breiman, L. (2001). Random forests. Machine Learning, 45(1): 5-32. https://doi.org/10.1023/A:1010933404324

[18] Juma, S., Muda, Z., Yassin, W. (2014). Reducing false alarm using hybrid intrusion detection based on X-Means clustering and random forest classification. Journal of Theoretical & Applied Information Technology, 68(2): 249-254. http://psasir.upm.edu.my/id/eprint/35184

[19] Moorthy, M., Sathyabama, S. (2011). Hybrid fuzzy

based intrusion detection system for wireless local area networks (HFIDS). Bonfring International Journal of Research in Communication Engineering, 1(Inaugural Special Issue): 26-30. https://doi.org/10.9756/BIJRCE.1006

[20] Abdelkader, A., Youcef, D., Hadjali, A. (2016). On the use of belief functions to improve high performance intrusion detection system. 2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Naples, Italy, pp. 266-270. https://doi.org/10.1109/SITIS.2016.50

[21] Ahmim, A., Derdour, M., Ferrag, M.A. (2018). An intrusion detection system based on combining probability predictions of a tree of classifiers. International Journal of Communication System, 28(12): 1428-1442. https://doi.org/10.1002/dac.3547

[22] The UCI KDD Archive Information and Computer Science University of California, Irvine. KDD Cup 1999 Data. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, accessed on 12 July 2018.

[23] Witten, I.H., Frank, E., Hall, M.A., Pal, C.J. (2016). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Burlington.

[24] Witten, I.H., Frank, E., Hall, M.A. (2005). Practical machine learning tools and techniques. Morgan Kaufmann, Burlington, 578.

[25] Cohen, W.W. (1995). Fast effective rule induction. In Proceedings of the Twelfth International Conference on Machine Learning, pp. 115-123.