COMPARISON OF HYBRID MACHINE LEARNING ALGORITHMS FOR INTRUDER DETECTION

[1]M. Chithik Raja, [2]M. Munir Ahmed Rabbani

Abstract

In the modern world Network information security has become an important concern of Internet and volume data. Web security services 'intrusion detection system (IDS) is a critical component that uses network traffic statistics to identify attacks. IDS should be able to implement data exploration and information mining systems to classify network machine outbreaks. The computation costs of IDS, however, is also necessary to help with the actual detection in line for the dismissal and inappropriate features in the Internet stream of traffic-dataset. We in this manner break down six Artificial Intelligence algorithm for IDS. Everyplace we independently execute information preprocessing with two sorts of dimensionality decrease procedures like Principal Component Analysis and Singular Value Decomposition to detect the attacks in the NSL-KDD dataset. The investigation results on the NSL-KDD dataset check that the course of action estimations with dimensionality drop out stands in discovery rate and speed. From our experimental result Principle Component Analysis with classifier KNN gives its matchless quality comparing with Singular Value Decomposition technique.

Keywords: *IDS, SVM, KNN, NSL-KDD*

I. Introduction

Intruder Detection System (IDS) is a system safety gadget that screens organize information in brisk and continues dynamic estimates when it seems far-fetched correspondences. Because of customary noxious system action and system approach infringement, IDS is comprehensively executed in different sorts of systems[1]. The essential issue of current IDS is that there are such an enormous number of qualities of framework information, and there may be a high association between specific properties, which makes the classifier unfit to exactly and quickly perceive the commonplace and atypical direct of the system [2]. Also, when IDS picks a subset of tests, it puts aside a lot of exertion to completely look and test each subset as a result of the element of the examples. Crushing information into relatively low-dimensional subspace is consequently of innumerable help[2]. Information dimensionality decrease diminishes the weights of putting away space, yet in addition accelerate the grouping algorithms. As a typical system for information preprocessing, dimensionality decrease, is applied to immaculate the clamor, and wrapping the information into a subspace of diminished measurement while holding the huge information to the most noteworthy degree. In any case, it may likewise diminish

[1] Department of Information Technology, Academy of Maritime Education and Training, Chennai-603112, India.
[2] Department of Computer Application, B. S. Abdur Rahman Crescent Institute Of Science And Technology, Chennai-600048, India.

Received: 02 Dec 2019 | Revised: 24 Jan 2020 | Accepted: 16 Feb 2020

the exactness of algorithms. Various exceptionally corresponded highlights, which are alluded to as excess highlights, nearby by highlights that tiny affect test grouping, which are alluded to as unessential highlights, causes a longstanding issue in organize traffic order. These highlights not just intentional down the procedure of arrangement and rise computational overhead, yet in addition stay away from a classifier from framing honest choices, especially when dealing with through huge information [3].

Disposal of dismissed as well as inconsequential component is the key objective in any component determination algorithm. In period of high-dimensional system traffic information, highlight determination can decrease the preparation time of the classification algorithm, diminish the computational cost of IDS, and along these lines illuminating the presentation. A similarly high identification rate as well as relatively high discovery speed are both fundamental factors for IDS. Finding the best ID algorithm isn't simple because of the absence of perfect preprocessing and order methods for recognizing anomalies. Ongoing advancements in IT have built up a wide assortment of AI models, which can be joined into IDS [4].

To improve the adequacy of IDS many regulated and unaided techniques from the field of deep learning and problem-solving perceptive have been conveyed. There are numerous regular classifiers, for example, Naive Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), and so forth. These classifiers have diverse characterization impacts for various datasets[5]. Classifiers without preprocessing have issues, for example, high computational cost as well as low recognition rate. Henceforth, we break down six classification systems: NB, LR, K-Nearest Neighbor (KNN), SVM, DT, AdaBoost (AB), and Random Forest (RF) in this paper. Head Component Analysis (PCA) and Singular Value Decomposition (SVD) are received to decrease computational overhead. Our work can be abridged as follows: diminishing the computational overhead with SVD and PCA[6]. contrasting six diverse order algorithms on pointers, for example, exactness, review, accuracy, and so on.; Analysis of the relating consequences of both the appropriation of PCA and SVD from the previously mentioned algorithms; investigation of the impact of measurement decrease.

The NSL-KDD Data set is proposed to resolve the two issues of the KDD-CUP-'99 Data set[7]. The major two issues are huge number of redundant records, which affect the effectiveness of evaluated systems greatly as a consequence and the prediction accuracy is unbelievably high. These issues lead to poor analysis result in intruder detection approaches[7]. To comprehend these issues, they proposed new Data set, known as NSL-KDD, which comprises selected records of the complete KDD-cup‘99 Data set. The KDDTrain+ and KDDTest+ sets of NSL-KDD Data set consist of 125,973 and 22,544 connection records respectively. Comparable NSL-KDD dataset, each record in this data set is unique with forty one features and labelled by means of regular or an attack [8]. NSL-KDD Data set contains the four types of attacks (Dos, Prope, R2L, U2R) . This set contains some new attacks that do not appear in the KDD Train+ data set which makes the detection of those attacks even harder[.9]. The bench mark Data set NSL-KDD is used to perform the experiment. There are many data sets are available today such as ISCX, Kyoto, CAIDA, NSL-KDD, UNSWNB15 etc. In this proposed work, NSL-KDD is the suitable data set, which is widely used and suitable for performance analysis. In this connection, we investigated comparison of hybrid machine learning algorithm for intruder detection.

II. Materials and Methods

Data Labelling

Data labelling because several features within the sample data are composed of letters, we might want to change over the qualities of the relating letters into numerical qualities to evacuate its impact on the algorithm. For eg, it consists of three kinds of data for the Protocol Type feature, namely Transmission control protocol, User datagram protocol, and Internet control message protocol. Since space process can't be done on these data, we are replacing them to ensure that these inaccessible features are available one by one with 0, 1, and 2, as shown in Table 1. For the type of test data labeling the conversion rule is: normal record is zero and abnormal record is one.

Table 1 Data Labelling of Protocol-Type in the feature

Actual-Eigen-value	Translated-Eigen-value
TCP	zero
UDP	one
ICMP	two

Maximum and Minimum Normalization

As the base estimation of specific feature inside the information is under 1 the most worth is a few thousands, which limits the utilization of distance based characterization algorithms, so we might want to standardize the ceaseless information. Min-max standardization is utilized here for normalization, as set out in (1). “Every stake feature is subtracted from the stake's minimum value, and also Minimum stake feature value. Where represents the standardized data, represents data, Min is the minimum value of each stake feature, Max is the most value of each stake feature”.

$$x_j^* = \frac{x_j - Min}{Max - Min} \qquad (1)$$

Dimensionality Reduction using Principle Comppnemt Analysis

PCA might be a measurable strategy for discovering designs in high-dimensional information. This proselytes segment related unique arbitrary vectors by a symmetrical change into new irregular vectors with uncorrelated parts. “N-dimensional high-request venture information to low-arrange k-dimensional information (n > k) without losing any significant delicacies. This transformation is implemented by the PCA by identifying k feature vectors, projecting n-dimensional data on the feature vector, thereby reducing the projection error [10]. According to Abielmona *et al.*, [10] illustrated in Fig. 1, Blue dots (represented by X1 and X2 features) are planned onto each of the 2 lines (Line 1 and Line 2). PCA Table 5.Protocol Type data mapping feature Original Ego-value Converted TCP=0,U=1and ICMP=3. Since the orthogonal projection error for projecting the information points onto Line 1 is far smaller compared to orthogonal projection error for projecting the identical data points onto Line 2”.

Received: 02 Dec 2019 | Revised: 24 Jan 2020 | Accepted: 16 Feb 2020

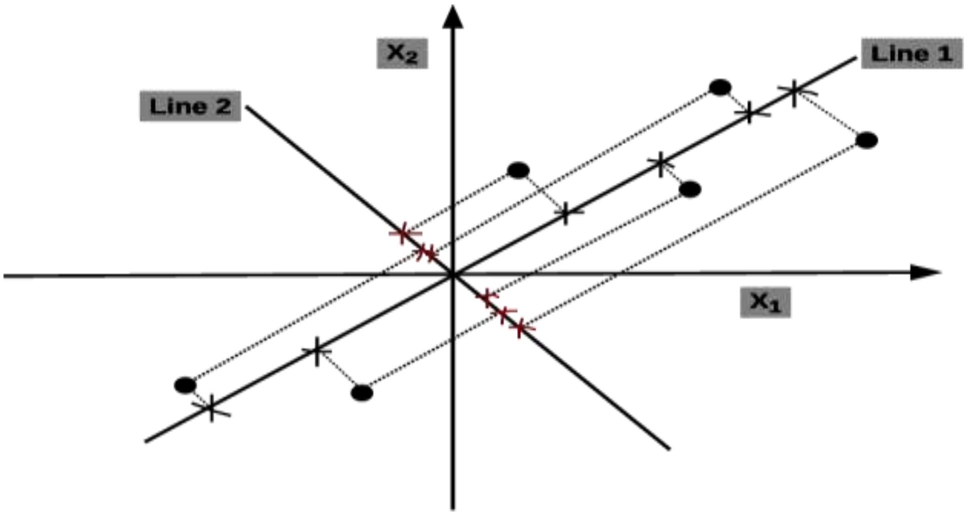

Figure - 1: 2-D representation of data arguments onto a 1-D line

Dimensionality Reduction using Single Value Decomposition

SVD might be a speculation of highlight decay on subjective grids. Expecting resultant matrix is m by n grid, at that point we characterize the SVD of matrix A as:

$$A = u\sum rv^{2t}(2)$$

According to Shuai and Xiaolong [11], investigated the matrix of m by m and V is an n by n matrix. R is an m by n matrix whose elements outside the most diagonal are all 0, and every element on the most diagonal is termed a singular value. Both U and V are unitary matrices, UTU = I, VTV = I. For singular values, it's the same as the eigenvalues in our feature decomposition. it's also arranged within the singular value matrix from large to small, and also the singular value is reduced especially fast. In many cases, the sum of the singular values of the primary 10% or perhaps 1% accounts for quite 99% of the sum of all singular values. That's to mention, we are able to also approximate the outline matrix with the biggest k singular values and also the corresponding left and right singular vectors (3).

$$A_{m\,X\,n} = U_{mXm} \ \Sigma_{mXn} \ V^T_{nXn} \ \approx \ U_{mXk} \ \Sigma_{kXk} \ V^T_{kXn} \qquad (3)$$

Where k is much smaller than n, a large matrix A can be represented by three small matrices[11.]

Experiment setup

The experiment setup was formulated according to Chandrashekar *et al* [21] and Rampure *et al* [22] employed NSL-KDD dataset for their model comparison using classifier algorithms. Based on their recommendation we use NSL-KDD open dataset, which tackles the inalienable issues inside the KDD-CUP-99 dataset. The training set of the NSL-KDD dataset doesn't contains repetitive records that the classifier doesn't predisposition to progressively visit records. Since the record number setting is moderate, this makes the investigation running on the total prearrangement of examinations cheap". Piuri *et al*. [14] have tentatively verified that NSL-KDD is that the best IDS records for classification designs. In spite of the fact that the NSL-KDD dataset consolidates few examples, it can in any case be utilized as a sound benchmark dataset, which may assist scientists with contrasting distinctive interruption recognition techniques. The records used for training eighty percent of the samples from the dataset, and furthermore the test set arbitrarily separates twenty percent of the samples from the dataset, which can keep the model from over-fitting and underfitting". It contains 67343 normal records as well as 58630 abnormal records [15.]

So as to encourage the correlation of succeeding execution, the accompanying markers are defined ahead of time. Classifier execution is assessed by ascertaining execution measurements, for example, Accuracy, Error rate, Precision, F-measure, AUC, and Detection time. According to Feng LiuJia "conditions (4) to (8), Where True positives (TP): Predicted positive and are actually positive. False positives (FP): Predicted positive and are actually negative. True negatives (TN): Predicted negative and are actually negative. False negatives (FN): Predicted negative and are actually positive. Accuracy: The most commonly used metric to judge. Error rate: the extent of mistakenly classified case; Recall/Detection rate: the extent of components effectively classified as positive out of every single positive component; Precision: the extent of components accurately classified as evident cautions out of the considerable number of components the interruption discovery model classified as positive; F-measure: It is the harmonic mean of precision and recall [25].

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

$$Error\ rate = \frac{FP+FN}{TP+TN+FP+FN} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

$$F-Measure = \frac{2*precision*recall}{precision+recall} \quad (8)$$

III. Results and Discussion

In this division, we assess the presentation of the model. All trials were performed on PCs running Windows-10 with IntelCorei7CPU @ 3.70 GHz and 16 GBRAM. To exhibit the preprocessing of information, we ran three investigations, individually: experiment 1 did not perform data preprocessing, and directly used six classification algorithms for classification; experiment 2 after PCA processing, using six classification algorithms for classification; experiment 3 after SVD processing, six classification algorithms are used for classification. Experiment 1 was carried out on 41 features of the original dataset, and Experiment 2 and Experiment 3 were performed on the first 23 features after dimensionality reduction.

Tables 2 and 3 show the exhibition examinations of the six Machine learning algorithms combined with PCA and SVD. When using six algorithms for classifying data without dimensionality reduction, RF, KNN, and DT are superior to other algorithms in Accuracy, Detection rate, Precision, F-measure, AUC, and Error rate. SVM, DT have less detection time than the other four algorithms.

Table 2 Comparision of six algorithm's result without using Dimensionality Reduction

	Accuracy	Recall	Precision	F-measure	AUC	Error rate	Time (ms)
KNN	0.997539	0.997607	0.997096	0.997352	0.997544	0.002461	7200

Received: 02 Dec 2019 | Revised: 24 Jan 2020 | Accepted: 16 Feb 2020

SVM	0.942965	0.896599	0.978823	0.935908	0.939888	0.057035	29
NB	0.898512	0.899932	0.883696	0.89174	0.898606	0.101488	175
DT	0.997619	0.998205	0.996672	0.997438	0.997658	0.002381	31
AB	0.980353	0.974022	0.983519	0.978747	0.979933	0.019647	900
RF	0.998412	0.997265	0.999315	0.998289	0.998336	0.001588	217

Table 3 Experimental Result in Hybrid of PCA with Six Machine Learning algorithms

	Accuracy	**Recall**	**Precision**	**F-measure**	**AUC**	**Error rate**	**Time (ms)**
KNN-PCA	0.996706	0.996753	0.996157	0.996455	0.996709	0.003294	5000
SVM-PCA	0.953126	0.940694	0.957634	0.949088	0.952301	0.046874	20
NB-PCA	0.900695	0.896257	0.890625	0.893432	0.9004	0.099305	80
DT-PCA	0.994999	0.994787	0.994447	0.994617	0.994985	0.005001	31
AB-PCA	0.977773	0.971629	0.980341	0.975966	0.977366	0.022227	741
RF-PCA	0.996507	0.994702	0.997771	0.996234	0.996387	0.003493	266

Especially in terms of computational overhead, the overhead of KNN and AdaBoost is too large, and KNN is about 2500 times that of SVM. The NB classifier has low indicators, and it is impossible to accurately perform intrusion detection for the current dataset. After the data of PCA or SVD dimensionality reduction is classified, the parameters of the six algorithms before and after the dimension reduction are not very different, and the time has been significantly improved. It can be seen that after data dimensionality reduction processing, although the data characteristics are reduced, it does not have an excessive negative impact on the accuracy of the classification and other indicators.

In addition, after the data preprocessing, the running time of the classifier is greatly reduced, and the average time consumption of KNN in PCA and SVD is 69.4% compared to without applying PCA and SVD. The reason why the six algorithms are greatly improved in time performance is that the feature dimension reduction greatly simplifies the dimension of the data and reduces the amount of data calculation during the detection process. Some algorithms have a slight improvement in each index because when the proposed model is applied, the obtained dataset cannot fully represent the original record, but the selected principal component contribution rate is over 95%. The redundancy has been cleared and the indicators have been improved.Their accuracy and other indicators are high compared to using all features. Some features in the original set of feature do not work to detect anomalies. The presence of these features will not only be a burden to detect anomalies, but will also increase the speed of notification.

The K-NN algorithms have longer training times and higher computational costs, while Singular Value Decomposition and Principle Component Analysis methods have the advantage of high computational speed and high operational efficiency, which can significantly reduce computational cost. By compare the combination of PCA with six classification algorithm and combination of SVD with six classification algorithms we obtained high detection and computational speed in hybrid of PCA and KNN algorithm. As shown in Fig. 2 PCA-KNN is greater accuracy compared to other algorithms. To reveal the performance of the hybrid technique PCA-KNN algorithm, we performed experiments using approximately 129,753 samples of the NSL-KDD data set.

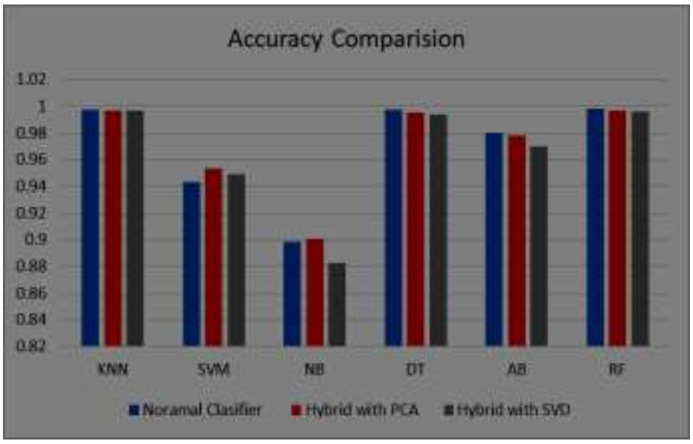

Figure 2 Accuracy Comparision of Six Algorithms with Normal and Hybrid PCA and Hybrid SVD

For this dataset PCA-KNN algorithm retain high accuracy with harder tracking time. Fig.3 shows the six algorithm's comparision result for all matrices. Table 5 shows result of six algorithm's comparision without dimentionality reduction. We found that RF algoristhm were given accuracy of 0.998412, recall value of 0.997265, F-Measure value of 0.999315, Precession value as 0.998289, AUC value of 0.998336 and Error rate as 0.001588 with execution time 217 miliseconds. RF algorithm is the best algorithm compared to all other algorithms when using without dimensionality reduction.Further we experimented with all six algorithms by using dimentionality reduction technique PCA and SVD. Table 6 shows the comparision result of six algorithm by using PCA technique.

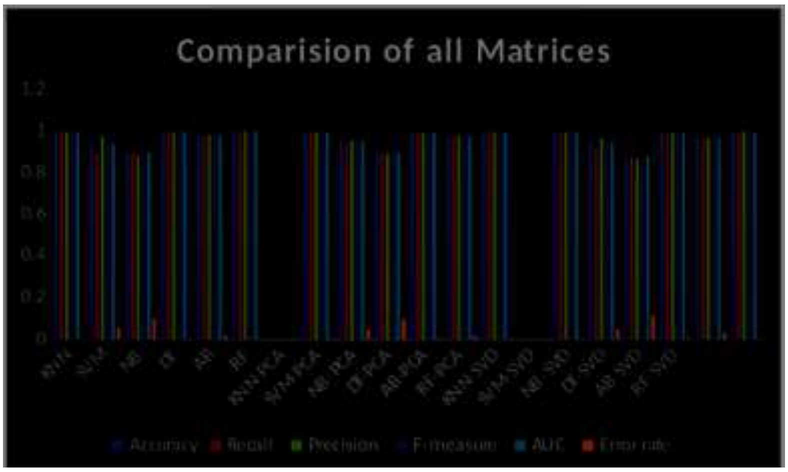

Figure - 3: All matrices comparision of six algorithms with Normal and Hybrid PCA and Hybrid SVD

Table 4 shows the comparision result of another dimensionality reduction technique SVD. While compared with all six alogithms by using PCA and SVD technique KNN were given good accuracy, Recall value and F-mesure value but its error rate and execution time were high compared to RF.   By using SVD technique RF algorithm were given less execution time very less error rate. We found that PCA-KNN were given best performance in terms of accuracy among six AI algorithms.

Received: 02 Dec 2019 | Revised: 24 Jan 2020 | Accepted: 16 Feb 2020

Table - 4 : Experimental Result in Hybrid of SVD with Six Machine Learning algorithms

	Accuracy	Recall	Precision	F-measure	AUC	Error rate	Time (ms)
KNN-SVD	0.996864	0.996838	0.996412	0.996625	0.996863	0.003136	5000
SVM-SVD	0.949355	0.922919	0.966529	0.944221	0.947601	0.050645	8
NB-SVD	0.882397	0.880619	0.868082	0.874305	0.882279	0.117603	99
DT-SVD	0.994086	0.994274	0.993002	0.993638	0.994099	0.005914	31
AB-SVD	0.969637	0.965647	0.968876	0.967259	0.969372	0.030363	802
RF-SVD	0.99611	0.994018	0.997599	0.995805	0.995971	0.00389	264

IV. Conclusion

In this paper we were experimented and compared the hybrid Machine learning algorithms. With respect to accuracy, precision and recall, an investigation of PCA with KNN findings is always greater. While compared with all six alogithms by using PCA and SVD technique in KNN were given good accuracy, Recall value and F-mesure value but its error rate and execution time were high compared to other algorithms. By using PCA technique KNN algorithm were given harder execution time very less error rate. As per our comparision result PCA with KNN is the best hybrid machine learning algorithm for the NSL-KDD dataset, because it reduces dimensionality by selecting appropriate feature vectors and discarding unimportant feature vectors. SVD and PCA can achieve an effective spatial reduction and redundancy elimination of data based on a maximum removal of a original data characteristics and a high computational overhead of the IDS. Concluded observations, we find that the PCA-KNN approach is superior in detection rate and time to other methods. PCA minimize the cost of computation and increase IDS performance while confirming Strong rate of detection. It has a certain meaning for the real-time deployment of high-speed networks. While PCA offers better performance analysis in terms of accuracy, error rate, recall, precision, F-measurement and AUC, according to complexities its needs a lot more computational time. Furthermore we can extend this approach for other modern dataset such as UNSW-NB15 which are present in the Canadian Institute for Cyber Security.

References

1. Meng W., Tischhauser E. W., Wang Q., Wang Y., Han J.: When intrusion detection meets blockchain technology: A Review. IEEE Access, vol. 6, pp. 10179–10188, (2018).
2. Kumar K., Singh J.: Network intrusion detection with feature selection techniques using machine-learning algorithms. International Journal of Computer Applications, vol. 150, pp. 1–13, (2016).
3. Ambusaidi M. A., He X., Nanda P., Tan Z.: Building an intrusion detection system using a filter-based feature selection algorithm. IEEE Transactions on Computers, vol. 65, pp. 2986–2998, (2016).
4. Aburomman A. A., Reaz M.B.I.: A survey of intrusion detection systems based on ensemble and hybrid classifiers. Computers & Security, vol. 65, pp. 135–152, (2017).
5. Ashfaq R. A. R., Wang X. Z., Huang J. Z.: Fuzziness based semi-supervised learning approach for intrusion detection system. Information Sciences. vol. 378, pp. 484–497, (2017).
6. Elejla O. E., Belaton B., Anbar M., Alnajjar A.: Intrusion detection systems of ICMPv6-based DDoS attacks. Neural Computing and Applications, vol. 30, pp. 45–56, (2016).
7. M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, “A Detailed Analysis of the KDD CUP 99 Data Set,” Submitted to Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2009.
8. Ikram S. T.: Improving accuracy of intrusion detection model using PCA and optimized SVM. Journal of Computing and Information Technology, vol. 24, pp. 133–148, (2016).
9. Subba B., Biswas S., Karmakar S.: Intrusion detection in mobile Ad-hoc networks: bayesian game formulation. Engineering Science & Technology, an International Journal, vol. 19, pp. 782–799, (2016).
10. Abielmona R.: 2012 IEEE Symposium on Computational Intelligence for Security and Defence Applications (IEEE CISDA 2012) [Conference Report]. IEEE Computational Intelligence Magazine, vol. 8, pp. 12–13, (2013).
11. Shuai Jiang and Xiaolong Xu , Application and Performance Analysis of Data Preprocessing for Intrusion Detection System vol. 20 pp. 163–177, (2019).
12. Chandrashekhar A. M., Raghuveer K.: Performance evaluation of data clustering techniques using KDD Cup-99 Intrusion detection data set. International Journal of Information and Network Security (IJINS), vol. 1, (2012).
13. Rampure V., Tiwari A.: A rough set based feature selection on KDD CUP 99 data set. International Journal of Database Theory and Application, vol. 8, pp. 149–156, (2015).
14. Piuri V. 2009 IEEE symposium series on computational intelligence (SSCI 2009) [Conference Reports. IEEE Computational Intelligence Magazine, vol. 4, pp. 20–21, (2009).
15. Ng K. Y. K., Lam M. N, The canadian forces’ information and intelligence fusion center: A preliminary capacity planning study1. Defense & Security Analysis, vol. 25, pp. 69–79 (2009).
16. Feng LiuJia XuShouhuai XuMoti Yung Science of Cyber Security, Second International Conference, SciSec 2019 Nanjing, China, August 9–11, (2019).

Received: 02 Dec 2019 | Revised: 24 Jan 2020 | Accepted: 16 Feb 2020