Implementation of Intrusion Detection System Using Modified K-Means Algorithm

Gulafshan[1], Prakash Mishra[2]

Computer science and Engineering Department,
Rajiv Gandhi Proudyogiki Vishwavidyalya, Bhopal, India.

Abstract: **The proposed work is motivated to implement an intrusion detection system using unsupervised learning technique. In order to design and develop a data mining based model the initial data samples or learning samples are required therefore the CIDS dataset is used for experimentation. This dataset is derived using the existing KDD CUP 99's dataset by refining the suitable and effective attributes. However the dataset contains of 21 attributes thus to reduce the dimensionality of dataset the correlation coefficient is used. Thus for each attribute the correlation coefficient is calculated and ranked according to the obtained values. The proposed technique is usage the modified k-means clustering algorithm for enhancing computational ability of classical algorithm. In this context the distance function of the k-means clustering is replaced with the RBF kernel function. The RBF kernel function is used here because the nature of data is not known initially. The modified kernel based k-means algorithm discovers the stable and centroids. Additionally by using the obtained centroids the test dataset is classified. Based on the classification outcomes the performance of the proposed system is measured in terms of accuracy. Additionally the resource consumption of the system is also noticed. According to the results the proposed model is efficient and accurate for classifying the binary labeled data.**

***Keywords*: Intrusion Detection system, k-means clustering, RBF Kernel, NSL –KDD Data set.**

I. INTRODUCTION

Now in these days a wide range of applications, individuals and organizations are become network and it's service consumers. The network may carry significantly confidential and sensitive information. On the other hand the network intruders and attackers are also become advanced and can deploy attacks on network from inside or outside of the network. In this context the intrusion detection systems are used for keep in track the security aspects. The intrusion detection systems are basically the software or hardware or the combination of both. That is used for evaluation of network traffic incoming to the inside or outside of network for finding abnormal behavior of network traffic. Therefore to identifying the unsolicited network behavior the IDS systems are much useful.

Therefore the proposed work is intended to design and efficient and accurate intrusion detection system using data mining techniques. Data mining techniques are basically used for analysis of data and pattern recovery. It supports various manners of pattern learning and recognition. In this work the unsupervised learning technique is tried to employ over the CIDS dataset for categorizing the network traffic into the malicious and normal network behaviors. However the dimension of input dataset is higher and can take a significant amount of time for processing the data. Therefore the feature selection technique is also applied for extracting the feature and optimizing the processing running time. Here the modified version of k-means clustering algorithm is used for classifying the intrusion patterns over the network traffic. To modify the k-means clustering algorithm the RBF (radial basis function) used. The RBF kernel is used in place of the distance function used traditionally with the k-means algorithm.

II. PROPOSED WORK

The proposed work is implementation and performance improvement of unsupervised learning based intrusion detection system. This section provides the details about the proposed data mining model additionally their functional aspects and their data processing methodology is also discussed.

A. System Overview

The data mining systems are accepted now in these days worldwide for analysis and pattern recovery in various applications among banking, engineering, medical and finance are the major domain. The data mining techniques are capable to automate the data capturing, refinement, and analysis therefore it become more advantageous. The data mining systems involve the various techniques and algorithms for processing data which is mainly categorize in two major classes supervised and unsupervised learning techniques. The supervised learning techniques are required initial examples to train the algorithms and after learning over the patterns or examples the algorithms are able to classify the data according to the experience. On the other hand the unsupervised learning techniques are usages the concept of data partitioning and optimization for recovering similar object based groups or categories. Therefore the unsupervised learning techniques are usages the different distance and/or similarity functions for evaluation of the data patterns.

In this presented work the unsupervised learning algorithm is applied on CIDS dataset for categorizing data patterns according to the class labels. The CIDS dataset is composed of two major class labels malicious and normal. The dataset is first prepossessed and then the feature is extracted. After obtaining the required features the improved k-means clustering algorithm is applied on data for categorizing according to the application requirements. For this purpose the RBF kernel function is applied with the algorithm for improvement. The section provides an overview of the proposed work involved and the next section provides the understanding about the functioning of thee proposed system.

B. Methodology

The proposed intrusion detection system is demonstrated using figure 2.1, this model define the input and the processing steps of the system using the components. There working and description is provided as:

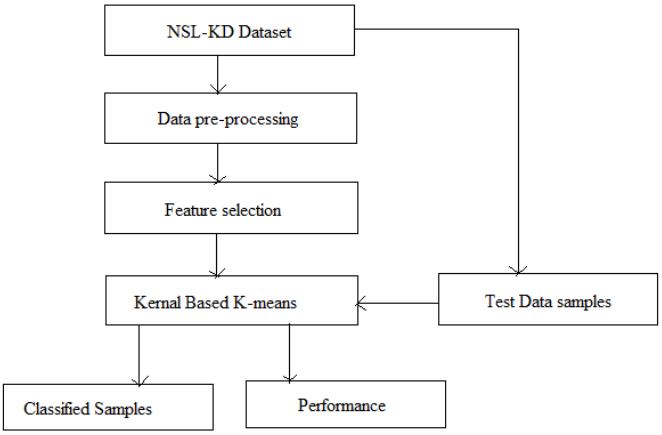

Figure 2.1 proposed system

NSL-KD Dataset: that is a popular bench mark dataset derived from the traditional KDD CUP dataset. Therefore irrelevant data attributes are removed from KDD CUP dataset for preparing this dataset. It contains total of 21 attributes means 20 is attributes and one is the class labels. Here the class labels are binary therefore 262,178 data instances are involved here as the attack pattern and 812,814 instance is available for normal traffic pattern. Using the entire dataset different size and composition based datasets are prepared.

Data preprocessing: the preprocessing of data in data mining is an essential step. Using this technique the data is refined for preparing it to utilize with the data mining algorithm. Therefore the data quality is improved in this step for effectively consume with the algorithm. In this work the missing values are removed during the preprocessing. Therefore all the data instances are located for removing the index instances. The below given process as given in table 2.1 can be used for preprocessing of data.

Input: dataset D Output: preprocessed data P
Process: 1. $[row, col] = readDataset(D)$ 2. $for(i = 1; i \leq row; i++)$ a. $for(j = 1; j \leq col; j++)$ i. $if(D_{i,j} == null)$ 1. $D.removeInstance(i)$ ii. End if b. End for c. $P.Add(D)$ 3. End for 4. Return P

Table 2.1 data preprocessing

Feature selection: the correlation is a statistical technique for measuring the bonding among two vectors. Therefore it is used here to find relevant attributes according to the distribution of class labels. Additionally by using a threshold value the matrix of dataset is reduced. Let we have a dataset with dimension m number of rows and n number of columns. Therefore, in terms of dataset vector format, m number of data instances and n number of attributes. So we need to reduce the x number of attributes to prepare the n-x size of dataset. In this context the correlation coefficient is computed among the available attributes and the class labels. The following equation is used:

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

Where N is the number of samples and x_i, y_i are the vectors which are need to be compared and

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Using this formula the correlation coefficient among all the attributes and class labels are computed and for reducing the dimensions of data a threshold value required. The following equation is used for threshold computation:

$$r_{threshold} = \frac{1}{N}\sum_{i=1}^{N} r_i$$

Thus the following algorithm is used for feature selection as given in table 2.2.

Input: input dataset D Output: Dataset Reduced dimensions R
Process: 1. $[row, col] = readDataset(D)$ 2. $y = D_{col-1}$ 3. $for(i = 1; i \leq col; i++)$ a. $temp = D_i$ b. $r_i = CoRc(y, temp)$ 4. End for 5. $r_{threshold} = \frac{1}{N}\sum_{i=1}^{N} r_i$ 6. $for(j = 1; j \leq n; j++)$ a. $if(r_i > r_{threshold})$ i. $R.Add(D_j)$ b. $end\ if$ 7. End for 8. Return R

Table 2.2 feature extraction

Kernel based k-means: traditionally the k-means clustering algorithm is a partition based clustering algorithm. That algorithm initially select k number of cluster centers randomly. After that the optimization process is taken place for finding the most suitable centroids. Therefore the main aim of the k-means clustering is to find optimal centroids by which remaining data instances can be defined. In this context the selected centroids are compared with the entire available data instances and based on the distance function their actual labels are assigned.

Traditionally for finding the distance Euclidean distance is used, if there are two vectors x_i and y_j then the distance among both the vectors can be defined using the following equation.

$$distance(x_i, y_j) = \sqrt{x_i{}^2 - y_j^2}$$

In addition of there are more functions available which can be used for finding the distance among two vectors. Therefore here the core distance function is replaced with the RBF kernel functions. The RBF kernel function can be described as:

$$k(x_i, y_j) = exp\left(\frac{1}{2\sigma^2}\|x_i - x_j\|^2\right)$$

The main aim of choosing this function is that we haven't any prior information about the nature of input data. Therefore RBF kernel is best suit for the nonlinear and unknown problem.

Test data samples: the proposed work is aimed to train the clustering algorithm by which the patterns are correctly recognizable. Therefore first the 70% of data instances are used for finding the most suitable cluster center. Additionally the 30% of samples are keep separated as test dataset for demonstrating the classification ability of elected centroids.

Classified samples: the test dataset is basically compared with the optimized centroids. The data instances which are closer to the optimal centroids are assigned to the class labels. Using this strategy the test dataset classified in two categories.

Performance: that is the final outcome of the proposed data mining system for intrusion detection. Based on the classified test samples the performance of the proposed IDS model is measured in terms of accuracy and error rate. Additionally with the increasing size of data instances the memory and time complexity is also measured.

C. Proposed Algorithm

This section provides the proposed algorithm steps to process and classify the data. The required process is given in table 2.3.

Input: Dataset D, number of cluster k = 2, Test Dataset T Output: classified instances C
Process: 1. $D_n = readDataset(D)$ 2. $P_n = preProcessData(D_n)$ 3. $[row, col] = GetSize(P_n)$ 4. $for(i = 1; i \leq col; i++)$ a. $R = getCorrelation(D_i, ClassLabel)$ 5. End for 6. $R_{threshold} = ComputeThreshold(R)$ 7. $for(j = 1; j \leq col; j++)$ a. $if(R_j \geq R_{threshold})$ i. $F.Add(D_j)$ b. $end\ if$ 8. End for 9. $cent = KKmeans.GetCentroid(F)$ 10. $T_m = ReadTestData(T)$

11. $for(k = 1; k \leq m; k++)$
    a. $if\ (dist(T_k, cent_{malicious}) \rightarrow 0)$
        i. $C = malicious$
    b. Else
        i. C = normal
    c. End if
12. End for
13. Return C

Table 3.3 proposed algorithm

III. RESULT ANALYSIS

The proposed work demonstrates the unsupervised learning technique for recovering malicious patterns in network traffic. In this section the capability of the system measured for finding the effectiveness of the proposed implemented algorithm.

A. Accuracy

Accuracy is a performance factor that is used to measure the correctness of a classification or categorization system. That is basically a factor that shows the amount of patterns is correctly identified during the algorithm execution. The accuracy can be calculated using this formula:

$$accuracy = \frac{total\ correctly\ identified\ samples}{total\ samples\ for\ classification} X100$$

Dataset Size	Accuracy (%)
1000	87
2000	91
3000	89
4000	88
5000	93
6000	91
7000	94

Table 5.1 accuracy in percentage (%)

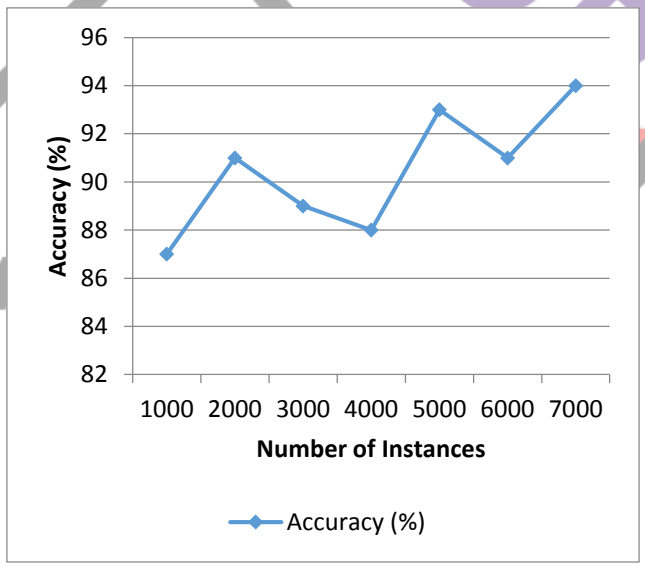

Figure 3.1 Accuracy in percentage (%)

The percentage accuracy of proposed kernel based k-means clustering algorithm for classification of malicious patterns is provided using figure 3.1 and table 3.1. The table 3.1 includes the observations of the experiments conducted on different size of datasets. The obtained values from the observations are represented in line graph as given in figure 3.1. As the results demonstrates the outcome of the proposed system is enhances with the increasing amount of data and influencing vary fewer with the noise available in the input dataset. Therefore the proposed technique is acceptable for the enhancing classification accuracy.

B. Error Rate

The error rate of a data mining system is the rate of misclassification or misrecognition. Therefore it is the measurement of error produced in classification. Thus, it is a ratio of misclassified data samples with respect to the total samples produced for classification. The measurement of error rate can be performed by using the formula described below:

$$error\ rate = \frac{Misclassified\ samples}{total\ samples\ to\ classify} X100$$

Or

$$error\ rate = 100 - percentage\ accuracy$$

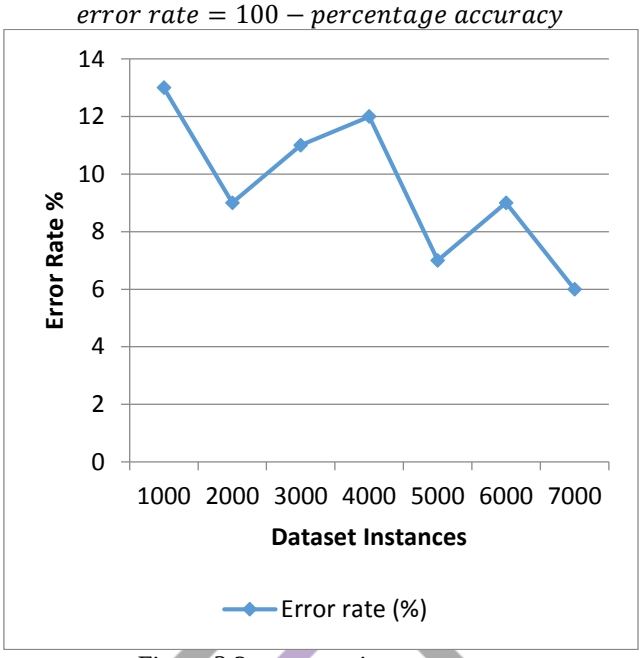

Figure 3.2 error rate in percentage

Dataset Size	Error rate (%)
1000	13
2000	9
3000	11
4000	12
5000	7
6000	9
7000	6

Table 3.2 error rate in percentage (%)

The experimental observation of measured error rate of the proposed technique for classifying malicious patterns is reported in table 3.2. Additionally their line graph representation is provided using figure 3.2. The computed error rate demonstrates the efficient outcomes of the proposed IDS (intrusion detection system). The results shows that the error rate is reduce with the amount of dataset instances are increases. Thus the proposed technique is acceptable in terms of error rate computation.

C. Memory Usages

The memory usages of the proposed intrusion detection system are described in table 3.3 and table 3.3. The memory usages of an algorithm are also known as the space complexity. That is computed as the amount of main memory acquired during the execution of an algorithm or process. For the java based implemented system it is computed using the following formula:

$$memory\ usages = total\ assigned - total\ free$$

Dataset Size	Memory usages (KB)
1000	22618
2000	23719
3000	22816
4000	22918
5000	23117
6000	23816
7000	23471

Table 3.3 memory usages

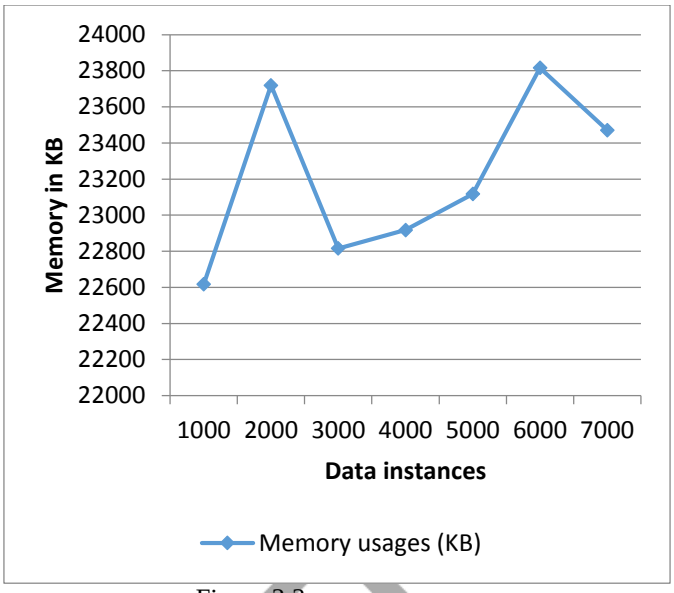

Figure 3.3 memory usages

The measurement of memory usages is performed in terms of KB (kilobytes). The figure 3.3 shows the line graph representation of the observed values from the table 3.3. The obtained line graph of the proposed system demonstrates the less resource consumption in terms of memory usages. Additionally the obtained memory usages of the system is not much fluctuating during the different experiments. Therefore the proposed kernel based k-means algorithm is demonstrating consistent performance for memory usages.

D. Time Consumed

That parameter is also known as time complexity of the system. The time consumption of the system is measured here for finding the total time consumed for processing the user input data using the implemented IDS system. The computation of time consumption is measured using this equation:

$$T_C = T_e - T_s$$

Where T_C is the consumed value of time, T_e is the end of processing time and T_s is the start time of algorithm.

Dataset Size	Milliseconds (MS)
1000	261
2000	519
3000	816
4000	1198
5000	1417
6000	1726
7000	2071

Table 3.4 time requirements of algorithm

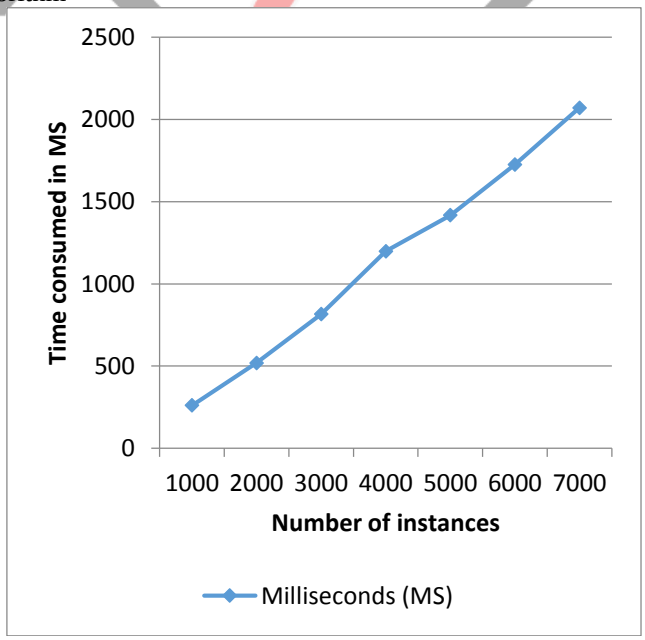

Figure 3.4 time consumed

The requirements of the time for processing the dataset using the proposed kernel based k-means are described in figure 3.4. The values of the line graph are reported using table 3.4. As shown in this diagram the time is increasing with the amount of dataset

instances. Therefore the time requirements of the data processing are dependent upon the amount of data. Here the X axis of diagram shows the number of instances in dataset and Y axis shows the consumed time in classification.

IV. CONCLUSION

The proposed work is motivated to design an improved intrusion detection system using unsupervised learning technique. In this context a system is developed successfully, this section provides the conclusion of the developed system. Additionally future extension is also suggested.

A. Conclusion

Rapidly expansion of the communication and need to network based system also motivate the attackers and intruders to perform the malicious activity. In this context the IDS (intrusion detection systems) are developed for monitoring of the network. The aim of monitoring is to evaluate the network traffic (i.e. internal and external) for finding the malicious patterns in network traffic. Using this method the intrusions of the system is identified. In this context data mining and machine learning techniques are become more fruitful for measuring and monitoring the network issues. Therefore this presented work is aimed to design and develop an efficient and accurate intrusion detections system.

The proposed intrusion detection system is developed using the unsupervised learning concept. Thus the k-means clustering algorithm is selected for the proposed system design. In this presented work for experimentation and system design CIDS dataset is used. That dataset contains the features of network traffic. But the data dimension is significantly higher therefore it is required to reduce the data dimension using some feature selection technique. In this context the correlation coefficient is used to rank the attributes according to relationship bonding with the available class labels. Therefore the higher ranked attributes in dataset is keep preserved and remaining attributes from the system is reduced. The less dimension data is efficiently processed by the proposed algorithm. In further a modified k-means algorithm is applied on data for categorizing the data into the number of available class labels. In order to modify the k-means clustering algorithm is used with the RBF (radial basis function). The RBF kernel is used here because we are not know about the nature of data and their relativity. Finally the dataset is classified and the performance of the system is measured and reported.

The proposed data mining system for IDS (intrusion detection system) design is developed using JAVA technology, additionally for preserving the measured performance of the system the MySql database is used. The measured performance parameters are with their observations and conclusion is reported in table 4.1.

S. No.	Parameters	Remark
1	Accuracy	The accuracy of the proposed model is found improved as compared to existing k-means algorithm for classification task
2	Error rate	The error rate of the proposed work is improved with the increasing amount of learning dataset
3	Memory usages	The memory usages of the proposed system is not much fluctuating and remains consistent
4	Time consumption	The time consumption of the proposed technique in a regular manner and directly depends upon the amount of data produced for categorization

Table 4.1 summary of performance

As reported in the above table 4.1 the performance of the proposed intrusion detection system is effective and able to reduce the false alarm rate. Therefore the proposed technique is acceptable for the future application development.

B. Future Work

The proposed work is to improve the existing IDS (intrusion detection system) is completed successfully. The implemented data mining algorithm able to classify the patterns accurately, therefore it is a promising algorithm. Additionally in near future the following work is proposed:

1. In this work the proposed data mining algorithm is modified with the RBF kernel, in near future the proposed work is extended with the other kernel functions
2. The proposed technique is an extension of existing method of unsupervised learning, in near future the proposed work is extended with a weight based clustering algorithm development.
3.

REFERENCES

[1] A. G. Karegowda, M. A. Jayaram, A. S. Manjunath, "Feature Subset Selection Problem using Wrapper Approach in Supervised Learning", ©2010 International Journal of Computer Applications (0975 – 8887), Volume 1 – No. 7
[2] S. Archana and Dr. K. Elangovan, "Survey of Classification Techniques in Data Mining", International Journal of Computer Science and Mobile Applications, Volume 2 Issue 2, February 2014.
[3] H. Jiawei, J. Pei, and M. Kamber, "Data mining: concepts and techniques", Elsevier, 2011.
[4] V. M. Saranya and Dr. S. Uma, "Survey on Classification Techniques Used in Data Mining and their Recent Advancements", International Journal of Science, Engineering and Technology Research, Volume 3, Issue 9, September 2014
[5] R. A. R. Ashfaq, X. Z. Wang, J. Z. Huang, H. Abbas, Y. L. He, "Fuzziness based semi-supervised learning approach for intrusion detection system", Information Sciences 000 (2016) 1–14
[6] C. Yin, Y. Zhu, J. Fei, and X. He, "A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks", VOLUME 5, 2017, 2169-3536, 2017 IEEE

[7] A. L. Buczak, and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection", IEEE Communications Surveys & Tutorials, Vol. 18, No. 2, Second Quarter 2016
[8] T. A. Tang, L. Mhamdi, D. McLernon, S. A. R. Zaidi and M. Ghogho, "Deep Learning Approach for Network Intrusion Detection in Software Defined Networking", 2016 International Conference on Wireless Networks and Mobile Communications (WINCOM), 26-29 Oct 2016
[9] P. Nader, P. Honeine, P. Beauseroy, "Detection of Cyberattacks In a Water Distribution System Using Machine Learning Techniques", ISBN: 978-1-4673-7504-7 ©2016 IEEE
[10] F. A. Narudin, A. Feizollah, N. B. Anuar, A. Gani, "Evaluation of machine learning classifiers for mobile malware detection", Soft Comput, DOI 10.1007/s00500-014-1511-6, Springer-Verlag Berlin Heidelberg 2014
[11] N. Sultana, N. Chilamkurti, W. Peng, R. Alhadad, "Survey on SDN based network intrusion detection system using machine learning approaches", Peer-to-Peer Networking and Applications, https://doi.org/10.1007/s12083-017-0630-0, part of Springer Nature 2018
[12] S. Aljawarneh, M. Aldwairi, M. B. Yassein, "Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model", Journal of Computational Science xxx (2017) xxx–xxx
[13] S. Agrawal, J. Agrawal, "Survey on Anomaly Detection using Data Mining Techniques", Procedia Computer Science 60 (2015) 708 – 713
[14] J. Kim, J. Kim, H. L. T. Thu, and H. Kim, "Long Short Term Memory Recurrent Neural Network Classifier for Intrusion Detection", International Conference on Platform Technology and Service, 2015