Improving Intrusion Detection System using PSO and SVM Algorithm

Shwetamaskare
Student, M.Tech
CS/ IT, Sage University, Indore, M.P., India

Shubhadubey
Assistant Professor
CS/ IT, Sage University, Indore, M.P., India

ABSTRACT
The new computational requirements are growing every day, and taken advantages of these services. But these networks are not fully secured a significant amount of attacks can be deployed on these networks. Therefore to secure the network from the attackers and malicious activities the proposed work is motivated to deliver enhanced IDS (intrusion detection system). That IDS is a data mining algorithm based technique for classifying the malicious patterns. In order to implement this technique the KDD CUP dataset is used. That dataset contains 41 attributes and 1 class attribute. This huge dimension can impact on the performance of IDS system. Therefore first the data processing technique is used to cleaning the data. After that the PSO (Particle swarm optimization) technique is used. Using this algorithm , rank all the attributes and select the features. The selected features are less in size means it contains 21 attributes and 1 class attribute. In this selected features the SVM algorithm is employed for classifying the data. The experimental results on different size of dataset demonstrate the effective performance of the proposed data model. That is also compared with the relevant k-NN classification model. The comparative performance analysis demonstrate the proposed model is accurate and less time consuming for classification of patterns as compared to the k-NN based model. But the memory usages of the proposed model are higher with respect to the k-NN model.

Keywords
IDS, data mining, PSO, SVM, classification, KDD CUP 99's

1. INTRODUCTION
Computer networks now in these days frequently used for communication channel, due to reachablity and accessibility of the public networks infrastructures. So, the computer network becomes much vulnerable because of the size of data communicated. Additionally it becomes more un-secure from the outside attackers as well as inside attackers [1]. The attacks deployed from inside of the network are known as internal attacks and the attack from outside of the network is known as the external attack. The IDS (intrusion detection systems) are helpful for identifying the attacks and dealing with intruders. That is an application to monitor network activity and user behaviors when users are actively utilizing network services. According to the activities of the network users and the functioning of network, the IDS make the decision for the attack [2].

Therefore in this work, an IDS is studied and the different available techniques are investigated. It is also recognizing how different researchers are improving the existing IDS systems. In conclusion, some of the improvements in the existing system are address. The improvement is focused on data mining based techniques employed with IDS. Here Data mining techniques are used to analyze the network traces using the algorithms and concluded the kinds of attack deployed. Moreover, making efforts to improving the detection accuracy of the IDS system of malicious activities in-network with low resource consumption. So, the work is intended to design and implement enhanced IDS, to learn on KDD CUP 99's dataset [3] and able to classify accurately. The basic need of the system is to reduce the false-negative rate. This section provides the overview of the proposed work involved in this study the next section provides the motivational article for the proposed work.

2. IMPROVED IDS SYSTEM
The main aim is, to improve the technique of current design of IDS for securing the network. In this context the proposed model which is a improved SVM system and their functional components are explained in this chapter.

2.1 System Overview
In recent years a significant amount of growth in digital network infrastructure is observed. These infrastructures enable us fast and low cost communication. Additionally provide the new applications and services for this context. Due to low cost and efficient networks not only individuals, large organizations and institutions are start storing the confidential information on digital storage and also accessing it frequently. These channels of communication are not much secured due to attackers and intruders. In order to prevent such kind of attackers a number of efforts are made in recent years among the IDS is one of the valuable technologies. The IDS help the network administrators to understand the network behavior and the user's activity who are consuming the network services. Thus, the IDS are an intelligent application which consume the network traffic pattern and provide the decision for network behavior normal or abnormal. In most of the designs the IDS systems implementing the techniques of machine learning and data mining which are much successful for such kind of complicated task.

In this work a supervised learning technique is used for designing the IDS system. To train this model samples obtained from the KDD CUP dataset were used. The training dataset contains the pre-classified samples with an identified class labels. The proposed algorithm consumes this dataset and train itself. In order to prepare the required data model two algorithms are employed namely the PSO (particle swarm optimization) algorithm and SVM (support vector machine). The PSO is used for feature extraction. That process helps to reduce time and space complexity due to reduction in the dataset dimensions. After computing the required features the SVM is applied for classification or to recognizing the target class labels. In this dataset we have five class labels/attacks (i.e. DOS, R2L, U2R, Probe, and normal). But, SVM is a binary classifier therefore training is performed as a "one Vs. all" technique. This section provides the basic overview of the proposed model next section details of the model is provided.

2.2 Methodology
The proposed machine learning based intrusion detection system is demonstrated in Fig.1. The components and their functional

aspects are explained in this section.

Input dataset: The initial step of the system is producing the dataset for learning into the system. Here, KDD CUP 99's dataset is used. That datasets are available in the UCI machine learning repository. That is a huge dataset that contains a total of 42 attributes among 1 attribute is class label as an example for the learning. In a supervised learning technique the first a training process required with some samples, using this training the model can identify the similar data pattern on which the algorithm trained. The dataset contains 41 attributes and 1 class label, so not all the information is in pure form. Thus, it contains a significant amount of information about the attack patterns. But it is an issue also relevant to the resource consumption. Additionally this data may contains various noisy content, and less informative attributes which are not required to be used for identifying an attack pattern. Therefore, the next process is used to optimize the data quality and reduction of dimension for achieving the accuracy and efficiency in terms of time and memory consumption.

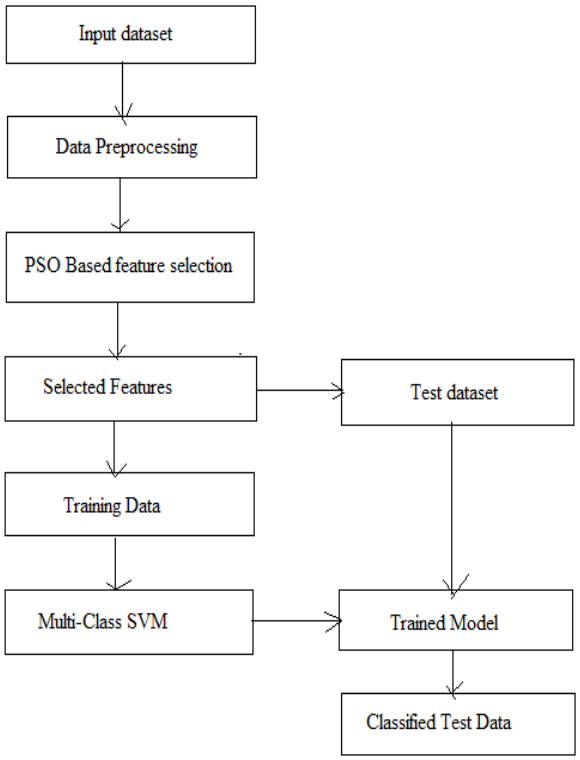

Fig 1: Working of Improved IDS SYSTEM

Data preprocessing: That technique basically used for cleaning of the dataset. This process is basically used for removing unwanted data from the dataset with different combinations of methods. There is not a fixed approach for data preprocessing. So, here used different techniques to reduce the noisy contents. In this presented work the null values or missing attributes are handled. In this context the input training samples are scanned for each data instance. During this scan the each pattern attributes are evaluated and when a missing value or null attribute found entire pattern is removed. Because it can influence the machine learning algorithm's decision making ability.

PSO based feature selection: After data pre-processing , get a new data with removal of those data instances which are incomplete. This process may reduce some amount of instance of data, so, the dataset size can be changed. This refined data is further used with PSO algorithm for selecting essential features. Thus first , the basic PSO algorithm is explained,

PSO (Particle Swarm Optimization) is introduced by Kennedy & Eberhart in 1995. That algorithm is inspired by the behavior and movement dynamics of insects, birds, and fish. That can recognized as Global gradient-less stochastic search method which is suited to continuous variables. PSO has successfully been applied to a wide variety of problems Neural Networks, Structural opt., Shape topology opt [35].

Advantages

- Insensitive to scaling of design variables
- Simple implementation
- Easily parallelized for concurrent processing
- Derivative free
- Very few algorithm parameters
- Very efficient global search algorithm

Disadvantages

- Slow convergence in refined search stage (weak local search ability)

The classical algorithm can be given as:

x_k^i = Particle position

v_k^i = Particle velocity

p_k^i = Best "remembered" individual particle position

p_k^g = Best "remembered" swarm position

c_1, c_2= Cognitive and social parameters

r_1, r_2= Random numbers between 0 and 1

Position of individual particles updated as follows:

$$x_{k+1}^i = x_k^i + v_{k+1}^i$$

With the velocity calculated as follows:

$$v_{k+1}^i = v_k^i + c_1 r_1 (p_k^i - x_k^i) + c_2 r_2 (p_k^g - x_k^i)$$

Algorithm Steps

1. Initialize
   a. Set constant k_{max}, c_1, c_2
   b. Randomly initialize particle position $x_0^i \in D$ in IR^n for $i = 0,1,2 \dots, p$
   c. Randomly initialize particle velocity $0 \le v_0^i \le v_0^{max}$ for $i = 0,1,2 \dots, p$
   d. Set $k = 1$
2. Optimize
   a. Evaluate function f_k^i using design space coordinates x_k^i
   b. If $f_k^i \le f_{best}^i$ then $f_{best}^i = f_k^i, p_k^i = x_k^i$
   c. If $f_k^i \le f_{best}^g$ then $f_{best}^g = f_k^i, p_k^g = x_k^i$
   d. If stopping condition reached then go to step 3
   e. Update all particle velocity v_k^i for $i = 1, \dots, p$
   f. Update all particle position x_k^i for $i = 1, \dots, p$
   g. Increment k
   h. Go to 2(a)
3. Terminate

Selected features: The PSO algorithm is an optimization technique. That technique is used here with the attribute basis for comparing the target values with the individual attribute and ranked all the attributes accordingly for selecting essential features among all the attributes. The higher ranked 21 features are selected here for experimentation.

Train set: The selected features are further going to be used for training and testing of the supervised learning algorithm. Therefore these features are subdivided into two subsets namely training and testing. The 70% of the instances from the selected features are random selected to be used for training.

Test set: In order to test the proposed model the remaining 30% of data samples are keep separately with their class labels used for testing. During validation these class labels are being used by algorithm for performance measurement.

Multi-class SVM: The SVM (support vector machine) is a supervised learning classifier. That is used for classifying the binary, but using various extensions we can use this as a multiclass classifier. The training dataset is used with a multi-class SVM. The SVM works here as a "one vs. all" concept for the learning. The trained Multi-Class SVM is further accepts the test dataset and recognizes the patterns of the test dataset.

Trained model: During the training of the SVM classifier it computes the various factors which can justify the current classification criteria. SVM is the most popular and useful techniques for data classification. It can be used for classifying linear and nonlinear datasets. The aim of SVM is to produce a model that predicts the target value of data that occurs in the testing in which only attributes have been given. The classification aim in SVM is to separate the two classes by a function prepared from dataset and producing a classifier that will further predicts classes for unseen data [10]. The basic form of SVM is the maximal margin classifier. It is used to solve the classification problems, namely the case of a binary classification with linear separable datasets. The aim of the maximal margin classifier is to explore the hyperplane with the largest margin, i.e., the maximal hyperplane; in real-world training data are not always linearly separable. In order to handle the nonlinear cases, some slack variables have been introduced to tolerate some training errors, with the influence of noise.

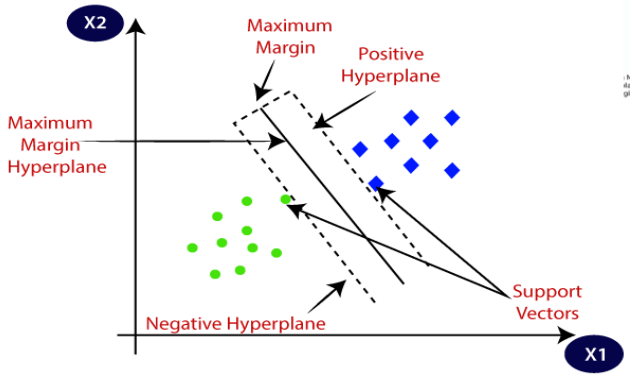

Fig.2 SVM Classification

This classifier with slack variables is referred to as a soft-margin classifier. The SVM classifier creates a hyper plane or multiple hyper-planes in high dimensional space useful for regression, classification, and other tasks. SVM has many features due to this it has gained popularity and has become promising with empirical performance. SVM constructs a hyper plane in reality.

Classified test data: That process helps in predicting the class labels of the input test instances. The classified test sample prediction of the performance of the system is measured. The measurement is taken in terms of accuracy, time, and space complexity.

2.3 Proposed Algorithm

The proposed intrusion detection technique is described here in terms of the algorithm steps. The steps are given in Table1.

Table 1 - Proposed Algorithm

Input: KDD CUP dataset D Output: Classified Samples C
Process: 1. $R_n = readDataset(D)$ 2. $P_m = PreprocessData(R_n)$ 3. $F_o = PSO.SelectFeatures(P_m, 21)$ 4. $[Train, Test] = F_o.Split(70,30)$ 5. $T_{model} = SVM.Train(Train)$ 6. $for(i = 1; i < Test.Length; i++)$ a. $C = T_{model}.Classify(Test_i)$ 7. $end\ for$ 8. Return C

3. RESULT ANALYSIS

The proposed work is motivated to design an improved intrusion detection system by using the techniques of data mining and machine learning. In this context a data model is implemented and their performance is evaluated and reported in this chapter.

3.1 Accuracy

The accuracy of a machine learning algorithm is an indicator of their correctness for identifying the patterns. The accuracy can be the fraction of the total correctly recognized patterns and the total patterns produced for recognition. That is calculated using the following formula:

$$accuracy(\%) = \frac{total\ correctly\ recognized}{total\ patterns} X\ 100$$

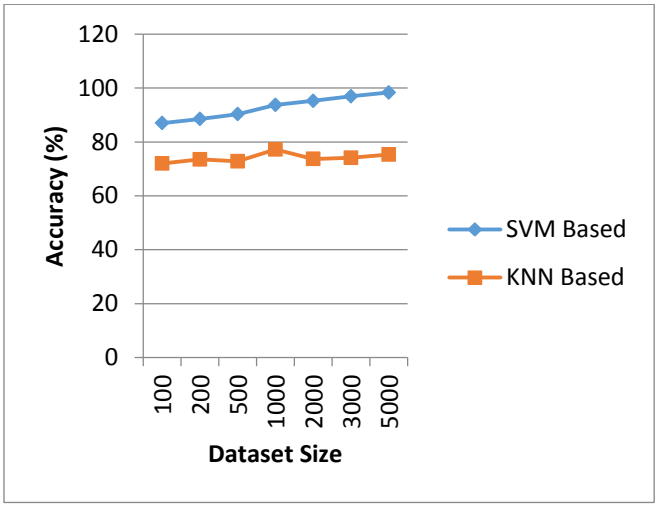

Fig.3 Accuracy (%)

Table 2 - Accuracy (%)

Dataset Size	Accuracy (%)	KNN Based
100	87	72
200	88.5	73.5
500	90.3	72.8
1000	93.7	77.2
2000	95.2	73.7
3000	96.9	74.1
5000	98.3	75.3

The performance of the proposed IDS system is demonstrated using Fig.3 and their observations are reported using Table 2. The Fig.3 visualizes the performance outcomes of the proposed model using line graph. In this line graph X axis contains the size of dataset used for experiments and the Y axis shows the performance obtained in experiments in terms of percentage (%). According to the gain results the proposed technique improves performance additionally with larger amount of data their performance become more suitable and accurate. The comparative performance of the proposed SVM based technique outperforms as compared to the classical KNN based classification model.

3.2 Error Rate

The error rate is basically providing information of misclassification rate during the classification. That is measured using the following formula:

$$error\ rate = \frac{Mis - classified\ Data\ Patterns}{total\ patterns} X100$$

Or

$$error\ rate\ (\%) = 100 - accuracy(\%)$$

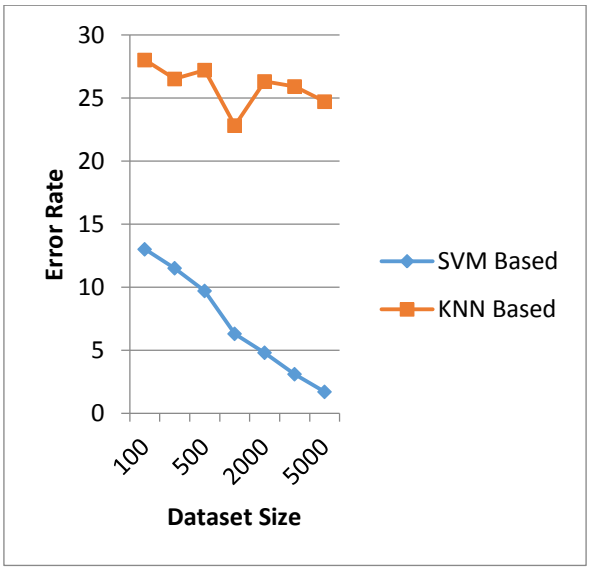

Fig.4 Error rate

Table 3 - Error rate

Dataset Size	SVM Based model	KNN Based
100	13	28
200	11.5	26.5
500	9.7	27.2
1000	6.3	22.8
2000	4.8	26.3
3000	3.1	25.9
5000	1.7	24.7

The Fig.4 shows the line graph of the obtained performance results. The error rate recorded is given in Table 3 and it's graphical representation is given in line graph. The X axis of this line graph shows the size of data used for experiments and the Y axis shows the recorded error rate. The error rate is measured here in terms of percentage (%). According to the obtained results the performance of model is become more accurate as the amount of data size is increases thus the model is acceptable for large amount of pattern classification in the security domain. The comparative performance of both the data models shows the SVM is much accurate classification technique as compared to the traditional k-NN and PSO based data model.

3.3 Memory Usages

The memory usages are also termed as the space complexity of an algorithm. The amount of main memory used during the algorithm processing is given here as the memory usages. In order to compute the memory usages of an algorithm in JAVA we can use the following formula:

$$memor\ usage = total\ assign\ memory - total\ free\ space$$

Table 4 - Memory usages

Dataset Size	SVM (KB)	K-NN (KB)
100	14262	14002
200	14628	14267
500	14992	14426
1000	15106	14718
2000	15400	14992
3000	15893	15217
5000	16291	15455

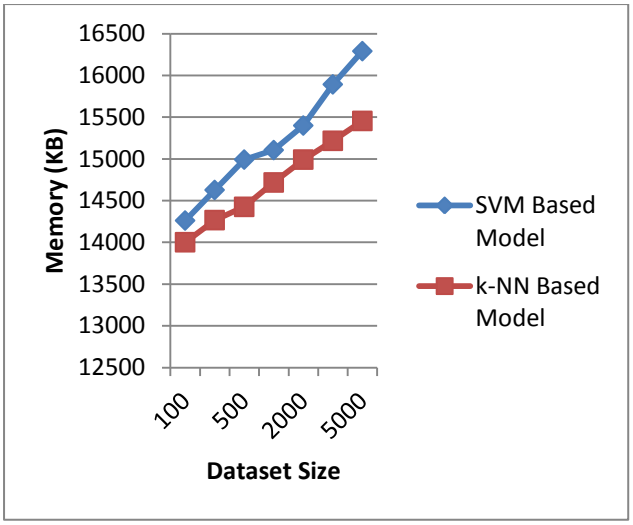

Fig.5 Memory Usages

The memory usages of the proposed data model are shown in Fig.5 and Table 4. The line graph as given in Fig.5 is graphical representation of values given in Table 4. The X axis of this line graph shows the dataset size used for experiments and Y axis of line graph shows the memory utilization of algorithm with respect to the dataset used during the experiments. The memory usages of the algorithm are measured here in terms of KB (kilobytes). According to the obtained performance the proposed model consumes less amount of memory but as the dataset size is increasing the amount of memory requirements are also increases in similar ratio. But the memory consumption of SVM based data model is higher as compared to k-NN based technique.

3.4 Time Consumption

Time requirement of an algorithm response is an essential factor of algorithm performance. That is also known as the time complexity of an algorithm. Here the time requirements of the algorithm are measured according to the size of data. To measure the time complexity of the algorithms the following formula is used:

$$time\ consumed = end\ time - start\ time$$

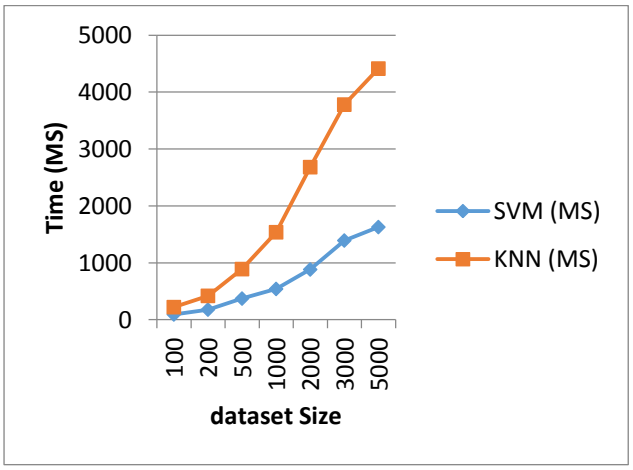

Fig.6 Time Consumption

The time requirement of the algorithm's training is given in Fig.6 and Table 5. The time is calculated here in terms of milliseconds (MS). The X axis of the given line graph shows the amount of data produces for training of the model in addition of Y axis of the line graph shows the corresponding time consumed for experiments. According to the size of data the time is increasing and decreasing. But the growth of time consumption is increases in the similar ratio as the amount of data produces as input. According to the comparative performance analysis the k-NN is a much time consuming algorithm. In the line graph we can see this difference much closely.

Table 5 - Time Consumption

Dataset Size	SVM (MS)	KNN (MS)
100	92	221
200	178	419
500	372	890
1000	541	1537
2000	883	2681
3000	1392	3780
5000	1628	4413

4. CONCLUSION

In this study a data mining technique based intrusion detection system (IDS) is implemented using the effective methods and algorithms. The observational and experimental facts are summarized in this chapter additionally the future extension of the work is also proposed.

4.1 Conclusion

Rapidly increasing digital infrastructures and applications invites the serious security issues in network. The existing solution for that context is not much effective and need improvements continuously. The rule based techniques are also not much effective due to each day the patterns of attacks may be different. Here the machine learning techniques are become valuable to train on complex patterns and recognize them more accurately. Thus, the proposed work is motivated to design an efficient and accurate data mining technique for classifying the malicious patterns available in IDS dataset. In this context the two major issues are we facing first the large amount of data for classification that directly impact on the performance of classification and the efficiency of the implemented programs. Therefore in this model two phases of dimensionality reduction is demonstrated first during the preprocessing the null values were removed and the second is use of PSO algorithm for selecting best features among all the dataset. The dataset contains a huge number of data instances and the significant amount of attributes i.e. 41 attributes and 1 class label. Thus the PSO is used for reducing the features for efficient classification. In order to apply the learning algorithm the SVM in one vs. all manner trained and their classification performance is recorded.

The proposed technique of IDS is implemented using the JAVA technology and with the help of Net Beans and MySql server. Additionally for classification algorithm implementation the WEKA library were used. After implementation of the required system the performance of the implemented model were evaluated and based on outcomes the performance summary is reported as given in table 4.1.

Table 6 - Performance Summary

S. No.	Parameters	Remark
1	Accuracy	The proposed model demonstrate higher accuracy and produces up to 98% of accurate classification rate
2	Error rate	The proposed model enhances their accuracy and reduces the error rate with the increasing amount of learning data that offers low error rate up to 1.7%
3	Memory usages	The memory usages the dependent factor of the amount of data used for experiment
4	Time consumed	The time consumption is acceptable and reducing gap of initial time consumption with increasing amount of dataset after learning

According to the performance obtained as reported in the Table 6 demonstrate the proposed model is effective model and can accurately recognize the malicious patterns of network. Thus the proposed model is used for real world IDS system design with small modifications.

4.2 Future Work

The aim of the proposed work of investigation of security and intrusion detection system is accomplished successfully in this work. Based on the collected experience the proposed model is effective and acceptable for offering security for networks IDS. The following extension is proposed for future design and implementation.

1. The model is currently utilizing the multi-class SVM in near future that is implemented using the CNN based techniques
2. The model include the PSO for implementing the feature selection technique in near future some more techniques are investigated for reducing the dimension of the dataset
3. The proposed work is needed to be extending for real time network traffic classification.

5. REFERENCES

[1] J. J. Jaccard, S. Nepal, "A survey of emerging threats in cybersecurity", Journal of Computer and System Sciences, 80, 2014, 973-993

[2] C. Modi, D. Patel, H. Patel, B. Borisaniya, A. Patel, M. Rajarajan, "A survey of intrusion detection techniques in Cloud", Journal of Network and Computer Applications, 36(1), pp. 42-57.

[3] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[4] M. H. Ali, B. A. D. A. Mohammad, A. Ismail, M. F. Zolkipli, "A New Intrusion Detection System Based on Fast Learning Network and Particle Swarm Optimization", VOLUME 6, 2018, 2169-3536, 2018 IEEE

[5] Wayne F. Cascio and Ramiro Montealegre, "How Technology Is Changing Work and Organizations", Annual Review of Organizational Psychology and Organizational Behavior March 2016

[6] M. Kashif, S. A. Malik, M. T. Abdullah, M. Umair, P. W. Khan, "A Systematic Review of Cyber Security and Classification of Attacks in Networks", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 6, 2018

[7] D. Denning, "An intrusion-detection model", Journal of Graph Theory, SE- 13(2): pp. 222–232, 1987.

[8] B. Mukherjee, L. Heberlein, and K. Levitt, "Network intrusion detection", Network, IEEE, 8(3): pp. 26–41, 1994.

[9] M. Joshi, "Classification, Clustering And Intrusion Detection System", International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 2,Mar-Apr 2012, pp.961-964

[10] V. Bande, U. D. Prasan, "Robust Intrusion Detection System using Layered Approach with Conditional Random Fields", IJCSET, October 2011, Volume 1, Issue 9, pp. 563-568

[11] F. Gorunescu, "Data Mining: Concepts, Models, and Techniques", Springer, 2011.

[12] J. Han, and M. Kamber, "Data mining: Concepts and techniques", Morgan-Kaufman Series of Data Management Systems San Diego: Academic Press, 2001.

[13] N. A. Padhy, Dr. P. Mishra and R. Panigrahi, "The Survey of Data Mining Applications and Feature Scope", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT)", vol.2, no.3, June 2012

[14] M. Rajalakshmi, M. Sakthi, "Max-Miner Algorithm Using Knowledge Discovery Process in Data Mining", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 11, November 2015

[15] "Data Mining Tutorial: Process, Techniques, Tools & Examples", available online at: https://www.guru99.com/data-mining-tutorial.html

[16] D. Denning, "An intrusion-detection model. Journal of Graph Theory", SE- 13(2): pp. 222–232, 1987.

[17] 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

[18] 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.

[19] H. Jiawei, J. Pei, and M. Kamber, "Data mining: concepts and techniques", Elsevier, 2011.

[20] 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

[21] H. S. Nair, S. E. V. Ewards, "A Study on Botnet Detection Techniques", International Journal of Scientific and

Research Publications, Volume 2, Issue 4, April 2012

[22] H. A. M. Uppal, M. Javed, and M. J. Arshad, "An overview of intrusion detection system (ids) along with its commonly used techniques and classifications", database, 19:20.

[23] V. M. Boncheva, "A short survey of intrusion detection systems", Problems of Engineering Cybernetics and Robotics, 58:23–30, 2007

[24] V. Engen, "Machine Learning for Network Based Intrusion Detection", June 2010, PhD. Dissertation, available online at: http://eprints.bournemouth.ac.uk/15899/1/Engen2010-PhD_single_sided.pdf

[25] M. H. Ali, B. A. D. A. Mohammad, A. Ismail, M. F. Zolkipli, "A New Intrusion Detection System Based on Fast Learning Network and Particle Swarm Optimization", Volume 6, 2018, 2169-3536, 2018 IEEE

[26] S. Balakrishnan, K. Venkatalakshmi, "Intrusion Detection System Using Feature Selection and Classification Technique", International Journal of Computer Science and Application, Volume 3 Issue 4, November 2014

[27] S. A. Mulay, P. R. Devale, "Intrusion Detection System using Support Vector Machine and Decision Tree", International Journal of Computer Applications, Volume 3 – No.3, June 2010

[28] Z. Dewa, L. A. Maglaras, "Data Mining and Intrusion Detection Systems", International Journal of Advanced Computer Science and Applications, Vol. 7, No. 1, 2016

[29] M. Srinivas, G. Janoski, A. Sung, "Intrusion detection using neural networks and support vector machines", Proceedings of the International Joint Conference on Neural Networks, IJCNN'02, Volume 2, IEEE, 2002.

[30] R. C. Chen, K. F. Cheng, "Using Rough Set and Support Vector Machine for Network Intrusion Detection", International Journal of Network Security & Its Applications, Volume 1, No 1, April 2009

[31] C. F. Tsai, C. Y. Lin. "A triangle area based nearest neighbors approach to intrusion detection", Pattern recognition 43.1 (2010): pp. 222-229.

[32] S. M. Othman, F. M. B. Alwi, N. T. Alsohybe, A. Y. A. Hashida, "Intrusion detection model using machine learning algorithm on Big Data environment", J Big Data (2018) 5:34, https://doi.org/10.1186/s40537-018-0145-4

[33] C. Yin, Y. Zhu, J. Fei, X. He, "A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks", Volume 5, 2017, 2169-3536, 2017 IEEE

[34] S. Das, A. M. Mahfouz, D. Venugopal, S. Shiva, "DDoS Intrusion Detection through Machine Learning Ensemble", 2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C), 978-1-7281-3925-8/19/$31.00 ©2019 IEEE

[35] J. F. Schutte, "The Particle Swarm Optimization Algorithm", EGM 6365 - Structural Optimization Fall 2005.