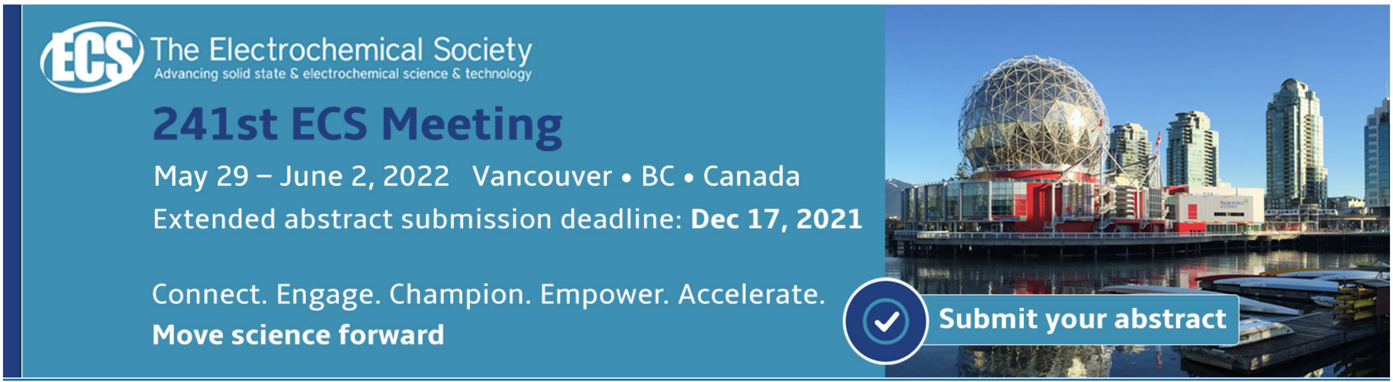
IOP Conference Series: Materials Science and Engineering

PAPER • OPEN ACCESS

Intrusion Detection System Using K-Means Based on Cuckoo Search Optimization

To cite this article: M. Deepa and Dr. P. Sumitra 2020 *IOP Conf. Ser.: Mater. Sci. Eng.* **993** 012049



INTRUSION DETECTION SYSTEM USING K-MEANS BASED ON CUCKOO SEARCH OPTIMIZATION

M.Deepa[1], Dr. P. Sumitra[2],
[1]Ph.D Research Scholar, Department of Computer Science,
legithasai2010@gmail.com
[2]Professor, Department of Computer Science,
drsumitra@vicas.org
[1,2]Vivekanandha College of Arts and Sciences for Women (Autonomous),
Elayampalayam

ABSTRACT

Recently, data protection is very important with the technological and digital revolution, as a vast amount of data is generated from different networks. It was found that the Intrusion Detection System (IDS) is probably the best option because of its ability to differentiate between threats that occur inside or outside a public internet. Cluster analysis is a common method of data mining and is characterized as the grouping of similar data. One of the clustering algorithms for clustering numerical data is K-Means. The K-Means Algorithm features are simple to implement and large amounts of data can be handled efficiently. Natural optimization algorithms have recently been combined with clustering algorithms in order to reach the best global solution. Algorithm for optimization search in Cuckoo is a recent meta algorithm for heuristic optimization. The intelligent behavior of the cuckoo is based on this algorithm. Cuckoo Search Optimization (CSO) and the K-Means clustering algorithm are combined in this paper to achieve the optimal solution globally. Different data sets are evaluated and results are compared with those of the clustering algorithms based on optimization.

Keywords: Intrusion detection system, k-means clustering, cuckoo search optimization**.**

I.INTRODUCTION

In recent years, the number of smart app users has grown exponentially. Network traffic has grown considerably. This extension has posed a host of safety concerns, including numerous possible or unexplained network assaults. Efforts and activities to jeopardise the security, integrity and/or functionality of a device or network can be identified as intrusions. IDS is one of the better techniques to track threats, as it involves a software or hardware device that tracks, measures and recognises continuous incidents both within and outside the network[1].

There are several ways of identification used by the IDS. The scheme or signature is compared with past incidents to identify existing threats in the identification of signatures. It is helpful to identify the threats already known, but it does not help to spot new threats, threats or hidden threats. Another method of detection is unusual detection, which compares the definition or characteristics of normal behaviour with the abnormality of the event[2]. Clustering is carried out in a variety of implementation fields, such as genetics, security, market analytics and social media-market analysis.

Clustering can be classified into two categories: hard and gentle. The same variable may only be part of a single cluster in hard clustering. During soft clustering, the same individual will belong to separate clusters. Cluster algorithms are categorised into two categories: partial and hierarchical. Clusters are classified into classes by dividing data structures into partition algorithms. Hierarchical clustering algorithms form clusters of data objects by hierarchical decomposition. K-means algorithm clusters are one of the classified algorithms and are general and commonly used because of their simplicity and efficiency[3,4].

Optimization algorithms have been developed on the basis of spontaneously inspired theories with the option of the right solution for such goals. Several optimization algorithms are Swarm-Based Algorithms (SBAs) and Evolutionary Algorithms (EAs). Crow Search Algorithm (CSA) and Artificial Bee Colony (ABC)[5][6] Particle Swarm Optimisation (PSO). Cuckoo Quest (CS) is also a type-inspired algorithm focused on the breeding technique of cuckoo birds for an increase in population. To keep fault frequencies and voltage variations within an acceptable range, a CS algorithm has been used to minimise individual power losses in a smart grid. Such breeding behaviour was idealised in the Cuckoo hunt and thus multiple optimization problems could be discussed. [8, 9] as indicated above.

A new clustering algorithm called CSOAKM is proposed in this article to solve the optimal local problem in K-Means clustering by using a Cuckoo Search Optimization and K-Means hybridization algorithm. The article shall be structured in the following order, the related work on the IDS shall be addressed in Section II and the solution proposed in Section III shall be discussed. The experiment is presented in Section IV. The Article shall conclude with section V.

II.RELATEDWORKS

This section discusses studies relevant to the proposed work. Overall, data mining algorithms, such as clustering algorithms, can be designed for anomaly-based IDSs. Data was originally obtained from the network flow or log data on the network. The clustering algorithm is then applied to these results. The clustering algorithm learns network data and generates a model. This model is referred to as an intrusion detection model. Identify the anomaly of packets that pass across the network or flow to the host.

Mistry suggested a genetic algorithm (GA) and a particulate swarm optimization (PSO) feature-selection method for facial emotion identification[10]. Ma et al suggested a wrapper-based approach to the collection of functions using the Knowledge Gain (IG) calculation with the help vector machine and machine learning algorithm to select relevant data set features to improve the precision of the classification [11]. The areas of intrusion detection were addressed with a decrease in dimensionality. The approach is to choose critical features for detecting new intrusions using the Ant Colony Optimization (ACO) algorithm and the nearest neighbour. Tests are conducted and measured on the NSLKDD dataset. 24 out of 41 features were chosen and 98.9 per cent accuracy and a false alarm rate of 2.59 per cent was achieved. The results suggest that the IDS can be successful with the proposed model [12].

The IDS performance is improved by the use of the information gain technique selection function and the use of SVM for Particle Swarm Optimization (PSO) classification. The results reveal that the 0.9 per cent false alarm rate and the 99.8 per cent sensitivity dominance of the IG-PSOSVM intrusion sensing model are extremely confidential in the NSL KDD dataset [13].

To boost the precision of the clustering model, researchers use the pre-processing methodology known as the selection of features or the selection of variables. Selection functionality is a process by which network data reduces redundancy and non-relevance. As a result, some investigators used tools for pre-processing results, such as the collection of functions to improve the precision of the IDS.

III. PROPOSED FRAMEWORK FOR IDS

A) K-MEANS ALGORITHM

K-means is one of the most unsupervised learning algorithms to overcome the well know clustering problem. This method is a simple and straightforward way to classify a particular data set in some fixed apriori clusters (assume k clusters). K centers, one for each cluster, are defined as the position with respect. Due to different locations, these centers should be placed in a smart manner. Therefore, it is easier to isolate them from each other as far as possible.

The next move is to make each data set point and link it to the nearest location. If no element is in progress, it completes the first stage and an early age is completed. We need k new centroids to be re-calculated as barycenters from the clusters resulting from the previous step. When we have these new centers, a new connection between the same data sets and the next new center is needed. This has built a loop. In this loop we will find that the K centers change their position step by step before changes are no longer made or, in other words, centers are no longer moving. Eq (1) measures this square error function.

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\left\|ai-bj\right\|\right)^{2} \qquad (1)$$

This is the definition of the K-Means clusteringalgorithm:
N number of objects is a_i.
Step 1: Enter the cluster as input.
Step 2: Choose the first K centroids randomly bj; j= 1; 2; ...; K from the data objects.

Step 3: Length between each center of the K-cluster and data objects are calculated using the formula

$$\sqrt{\sum_{i=1}^{n}\left(a_i - b_j\right)^2} \qquad (2)$$

Step 4: Search for the minimum distance and allocate cluster data objects.
Step 5: Update the centroids using the following equation

$$b_j = \frac{1}{N}\sum_{i=1}^{n} a_i \qquad (3)$$

The K-Means algorithm is terminated when one of the following conditions is satisfied:
(i) if the centroids' average change
(ii) if the iteration is reached the maximum limit and
(iii) The cluster objects is ideal.

B) CUCKOO SEARCH OPTIMIZATION

Cuckoo Search optimization (CSO) is a meta - heuristic optimization algorithm enhanced by the performance of the nature's brute parasitism. These birds never build their own nests and put eggs into other host nests for the expansion of their egg hatches [14,15]. This method is improved by levy flights instead of random measures to locate the nest. The great truths of these birds are that the host bird lays eggs and the birds impersonate the appearance of the host species. If the host learns that the egg is not theirs, they throw the eggs or they leave the house. Each house has an egg to overcome the honesty.

The CSO, which is meta-heuristic, depends on three ideal rules:
i) Each flying cuckoo lays an eggs at any time and chooses an odd location.
ii) In the new age, the best home with the highest egg quality is transferred.
iii) The host has quantity settled and the host is likely to find cuckoo egg.

STEPS FOR CUCKOO SEARCH ALGORITHM

Step 1: Specify the host nest initial value of size n.
Step 2: Generate the initial nest host population with optimum parameters (a_i, b_i).
Step 3: place the egg (ak',bk')in the K nest
Step 4: evaluate the fitness value of cuckoo's egg with the host's egg.
Step 5: if the fitness of cuckoo's egg is better than the host's egg then replaces the egg by cuckoo's egg
Step 6: if host bird notice it, the nest is abandoned and new one is built
Step 7: repeat step 2 to step 6 until termination criterion satisfied

C) PROPOSED ALGORITHM

The only drawback is that it lacks insufficient local solutions. K-Means is combined with global optimization algorithms to achieve the global optimum solution. In combination with KMeans, CSO is a global metaheuristicoptimisation algorithm for the best global solution. CSO and K-Means Algorithms are introduced in this section. The CSOAKM idea implies the following algorithm:

1. Enter the number of clusters as input k, host nest initial value n.
2. Generate the matrix of size K*N with random numbers (features in the dataset). The maximum range of values is equal to the the total number of instances in the data sets.
3. Compare the length of centroids and fitness values of egg initialize the memory of the crows with the values of the positions of the crows because initially crows hid their foods at their initial positions.
4. Evaluate the fitness of initial position of eggs.
5. If the cuckoo's fitness is better than the host than replace the host egg
6. Repeat the above steps until reach the optimum solution

IV. EXPERIMENTAL SETUP AND RESULTS

This research examines the findings and results of the success of the proposed system for detecting anomaly intrusions. A large number of irrelevant and repetitive features result in poor training and high running time. Thus, in order to resolve this problem, a large number of forms of IDSs have been proposed using various collection techniques. This research proposed a new feature range, which provides a powerful anomaly intrusion detection system, which will allow for better IDS efficiency and time reduction.

A. EXPERIMENTAL SETUP

Python is conducting the experiment with the Inter Core i7 2.7GHZ processor, the RAM is 8 GB and Eclipse and Anaconda 2.7 SCIkit-learn are the main software framework for the research. The collection of features is to delete information which is obsolete and meaningless. The CSOAKM solution is proposed to address this issue. The output is evaluated on NSL-KDD dataset taken from the UCI learning repository [16]. For 2, 5 and 10 fold cross validations; the proposed solution is tested for specific k values to assess broader outcomes. If the k-value is high, the training samples may be divided in k samples with small tests, the variability can be decreased, and if k is low, training and test samples are divided into k samples that have a large test range. The number of host nests (n population size) and the frequency of discovery Pa have been attempted to differ. For n(5,10,15,15,20,25,0.30,50, and Pα(0,0.01,0.05,0.0,0.1,0,21,0.15,0.2,0.25,0.4,0.5) we have used different settings. We found that n = 30 and Pα = 0,25 were adequate for most optimization problems from test phase simulations. We used fixed n = 30 and Pα = 0.25 to achieve the optimum by cuckoo search method and maximum number of iterations 100.

B. DATASET NSL-KDD

The efficacy of the proposed CSOA feature-selection anomaly detection and K-means clustering was investigated and compared to several other existing methods in the same field. The NSL-KDD is used to verify the dataset. It's for an offline IDS test. Table I displays the features in the NSL-KDD dataset. The data set contains 43 features per text, 41 of which apply to traffic input, and the last two features are labels (whether an attack or a normal one) and ratings (intensity of traffic input itself) and four attack types are probe, DoS, U2R, and R2L[17][18]. The NSL-KDD data collection is then designed to boost the KDD Cup 99 data and resolve the particular issues of the last mentioned.

The dataset is extracted from the separate parts of the original KDD Cup 99 dataset, with no unnecessary elements or repetitions. In addition, the issue of imbalanced dissemination in the research and training set has been overcome in order to increase the consistency of the IDS evaluation. The traffic record functions include the traffic entry information from the IDS and can be divided into four categories: Intrinsic, Material, Host- and Time-based. Intrinsic features can be extracted without looking into the payload itself from the header of the packet and contain the basic packet information. The characteristics of this group are 1–9. Content features include information on the initial packets as it is sent in more than one piece. The device can access the payload with this knowledge. The features in this group are 10–22.Time-based features include a two-second window review of the traffic data and information on the numbers of connections it has attempted to connect to the same host. Such characteristics are primarily counts and levels rather than information on the traffic input material. This group contains 23–31 characteristics.Host-based features are similar to timescales but to measure how many requests have been made to the same Server over x-number connections instead of measuring a 2-second window. Such characteristics are designed to accelerate attacks that stretch over a time span of two seconds. Attributes in this group are 32–41.

Table 1: NSL-KDD Dataset Features and its Descriptions

Feature name	**Description**
Duration	Connection Length.
Protocol Type	Protocol name
Service	Network service
Flag	Normal or Error
Src Bytes	Data bytes transferred from source to destination
Dst Bytes	Data bytes transferred from destination to source
Land	0-> if source and destination IP are similar. 1->otherwise
Wrong Fragment	Number of wrong fragments
Urgent	Number of urgent packets
Hot	Number of "hot" indicators
Num Failed Logins	Failed login attempts
Logged In	Login Status : 1 if successfully logged in; 0 otherwise
Num Compromised	Number of "compromised" conditions
Root Shell	1 if root shell is obtained; 0 otherwise
Su Attempted	1 if "su root" command attempted or used; 0 otherwise
Num Root	Number of "root" accesses
Num File Creations	Number of file creation
Num Shells	Number of shell prompts
Num Access Files	Number of operations on access control files
Num Outbound Cmds	Number of outbound commands in an ftp session
Is Hot Logins	1 if the username belongs to the "hot" set, i.e. root or admin, or 0

Is Guest Login	1 if the username is a 'guest' username; 0 otherwise.
Count	Amount of links to the same destination host as the existing link in the last two seconds.
Srv Count	Amount of links to the same infrastructure (port number) as the current connexion in the last two seconds.
Serror Rate	The percentage of connexions that enabled flag (4) s0, s1, s2 or s3 between connexions aggregated in count (23)
SrvSerror Rate	The percentage of connexions that have been allowed by flag (4) s0, s1 , s2 or s3, among the connexions aggregate d by srv count (24)
Rerror Rate	The percentage of connexions that enabled the flag (4) of the REJ between the connexions aggregated in count (23)
SrvRerror Rate	The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in srv_count (24)
Same Srv Rate	The percentage of connections that were to the same service, among the connections aggregated in count (23)
Diff Srv Rate	The percentage of connections that were to different services, among the connections aggregated in count (23)
Srv Diff Host Rate	The percentage of connections that were to different destination machines among the connections aggregated in srv_count (24)
Dst Host Count	Number of connections having the same destination host IP address
Dst Host Srv Count	Number of connections having the same port number
Dst Host Same Srv Rate	The percentage of connections that were to different services, among the connections aggregated in dst_host_count (32)
Dst Host Diff Srv Rate	The percentage of connections that were to different services, among the connections aggregated in dst_host_count (32)
Dst Host Same Src Port Rate	The percentage of connections that were to the same source port, among the connections aggregated in dst_host_srv_count (33)
Dst Host Srv Diff Host Rate	The percentage of connections that were to different destination machines, among the connections aggregated in dst_host_srv_count (33)
Dst Host Serror Rate	The percentage of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in dst_host_count (32)
Dst Host SrvSerror Rate	The percent of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in dst_host_srv_count (33)
Dst Host Rerror Rate	The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in dst_host_count (32)
Dst Host SrvRerror Rate	The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in dst_host_srv_count (33)
Label	Classification of the traffic input
Score	Number of attacks

C) PERFORMANCE ANALYSIS

The current feature selection algorithms, namely Information Gain-Naïve Bayes (IG-NB), chi square selection, andCuckoo optimization forfeature selection (COFS) [19] are used in comparing output with the proposed method. Moreover, the K-Means clustering algorithm is employed to build the intrusion detection model and evaluate the performance of the feature-selection methods.

Table 2: Performance comparison for IDS methods

Number of	**COFS**	**IG-NB**	**Chisquare**	**Proposed IDS**

instancing				**CSOAKM**
2519	96.880	62.616	53.270	90.63
3779	80.750	62.305	62.616	92.02
5038	75.700	62.813	62.305	94.16
7557	69.862	64.043	63.470	97.16
10000	63.470	69.470	66.043	99.8

The experiment shall be performed as follows: Initially, each feature selection method is given the data set and the k value (number of characteristics to be selected), while k is a selection method. Then, k number of features selected by the feature-selection method along with the class attribute is given to the K-Means clustering algorithm, and the accuracy is calculated as tabulated in Table 2 to determine the performance of the featureselection methods.

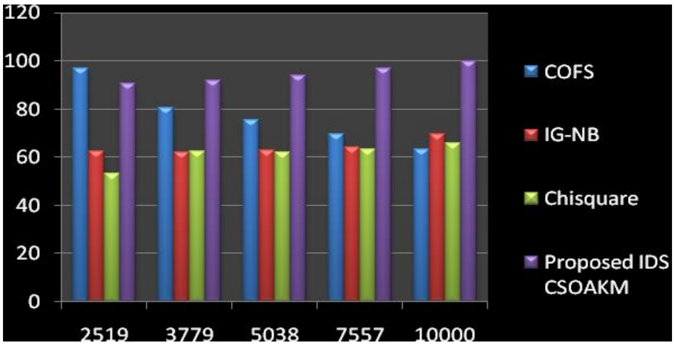


Figure 2: Accuracy comparison of various IDS methods

The evaluation of the data set to detect attacks concerns 41 features. Among the 41 attributes the COFS algorithm select 25 features, chi square algorithm select 33 features, CSOAKM selects 19 features. The performance is evaluated based on 19 features. From Table 2, Figure 1, it is found that in comparison to other approaches, the proposed approach produces greater precision for the model of intrusion detection.

V. CONCLUSION AND SUMMARY

The analysis is structured to select the best feature subsets and the K-means algorithm for clustering purposes using the latest feature selection called "Cuckoo Search Optimization" (CSO). This model has been applied to the problem of intrusion sensing and is validated using the well-known NSL-KDD dataset. The selection process included 19 of the 43 features that are key to increasing output. In addition, this model compares existing models in a similar experimental setting. The proposed CSOAKM exceeds the current models in terms of identification, false alarm rate and execution time, and the results suggest that the proposed model could be appropriate for the IDS. In order to detect attack type, the IDS model can be extended to cover problems of multi-class ranking, further efforts will be made to improve classification accuracy. In addition, we will test the proposed model for future research using the data set for intrusion detection in real time.

VI. REFERENCES

[1] Ali,Mohammed, "A new intrusion detection system based on Fast Learning Network and Particle swarm optimization," IEEE Access, vol. XX, no. c, pp. 1–1, 2018.

[2] Yi Aung, Min , Hybrid Intrusion Detection System using K-means and K-Nearest Neighbors Algorithms , IEEE ICIS 2018, June 6-8, 2018, Singapore

[3] Han J, Pei J and Kamber M 2011 Data mining: concepts and techniques. Elsevier, United States

[4] VISALAKSHI, SHANTHI,ETC, Data clustering using K-Means based on Crow Search Algorithm, Sådhanå (2018) 43:190 Indian Academy of Sciences https://doi.org/10.1007/s12046-018-0962-3Sadhana(0123456789().,-volV)FT3](0123456789().,-volV)

[5] M. Shehab, A. T. Khader, and M. A. Al-Betar, "A survey on applications and variants of the cuckoo search algorithm," Appl. Soft Comput.J., vol. 61, pp. 1041–1059, 2017.

[6] Samira,etc, an efficient anomaly intrusion detection method with feature selection and evolutionary neural network, citation information: doi 10.1109/access.2020.2986217, ieee access.

[7]R. Priyadharshini, E. Jebamalar, Cuckoo Optimisation based Intrusion Detection System for Cloud Computing, I. J. Computer Network and Information Security, 2018, 11, 42-49 Published Online November 2018 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijcnis.2018.11.05

[8]] W. Buaklee, K. Hongesombut, Optimal DG allocation in a smart distribution grid using Cuckoo search algorithm, ECTI Trans. Electr. Eng. Electron.Commun. 11 (2) (2013) 16–22.

[9] https://en.wikipedia.org/wiki/Cuckoo_search

[10] J. H. Mohamud and O. N. Gerek, "Poverty level characterization via feature selection and machine learning," 27th Signal Process. Commun.Appl. Conf. SIU 2019, pp. 1–4, 2019.

[11]Mistry K, Zhang L, Neoh SC, Lim CP, Fielding B. A micro-GA embedded PSO feature selection approach to intelligent facial emotion recognition. IEEE Transactions on Cybernetics 2016; 47(6):1496–509.

[12] Ma L, Li M, Gao Y, Chen T, Ma X, Qu L. A novel wrapper approach for feature selection in object-based image classification using polygon-based cross-validation. IEEE Geoscience and Remote Sensing Letters 2017; 14(3):409–13

[13] M. H. Aghdam and P. Kabiri, “Feature selection for intrusion detection system using ant colony optimization,” Int. J. Netw.Secur., vol. 18, no. 3, pp. 420–432, 2016.

[14] E. M. Chakir, M. Moughit, and Y. I. Khamlichi, “An effective intrusion detection model based on svm with feature selection and parameters optimization,” J. Theor. Appl. Inf. Technol., vol. 96, no. 12, pp. 3873–3885, 2018.

[15].M. Lichman, UCI Machine Learning Repository, 2013, http://archive.ics.uci.edu/ml.

[16] C. Chio, “Machine learning based techniques for network intrusion detection,” HackInParis, pp. 79– 83, 2016.

[17] A. G. M. Tavallaee, E. Bagheri, W. Lu, “Canadian Institute for Cybersecurity.” [Online].

[18]. H. Hindy et al., “A Taxonomy and Survey of Intrusion Detection System Design Techniques, Network Threats and Datasets,” vol. 1, no. 1, 2018.

[19]. Frank E, Hall MA, Witten IH. The WEKA workbench. In: Kaufmann M, editor. Online appendix for data mining: practical machine learning tools and techniques. 4th ed. 2016

[20].Singh G, Antony DA, Leavline EJ. Data mining in network security-techniques and tools: a research perspective. Journal of Theoretical and Applied Information Technology 2013; 57(2):269–78.