2017 International Conference on Trends in Electronics and Informatics (ICEI 2017)

Tirunelveli, India
11-12 May 2017

Pages 584-1180

IEEE Catalog Number: CFP17J32-POD
ISBN: 978-1-5090-4258-6

IEEE Catalog Number:	CFP17J32-POD
ISBN (Print-On-Demand):	978-1-5090-4258-6
ISBN (Online):	978-1-5090-4257-9

Additional Copies of This Publication Are Available From:

Curran Associates, Inc
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: (845) 758-0400
Fax: (845) 758-2633
E-mail: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

Feature Selection Based Intrusion Detection System Using the Combination of DBSCAN, K-Mean++ and SMO algorithms

Vandana Shakya
Department of CSE/IT
Madhav Institute of Technology & Science
Gwalior (M.P.) India
vandana.shakya12@gmail.com

Rajni Ranjan Singh Makwana
Department of CSE/IT
Madhav Institute of Technology & Science
Gwalior (M.P.) India
rrsingh@mitsgwalior.in

Abstract—IDS is the main concern of the security which is useful to prevent the attack at host and network level. In this propose work, classification of KDD intrusion dataset is proposed along with noise reduction, clustering and feature selection. DBSCAN algorithm has been applied to reduce noise present in KDD dataset. After noise removal genetic search approach is utilize to pick relevant feature. K-Means++ clustering method is utilized to cluster the dataset and resultant dataset is tested by SMO based classifier. A confirmation of concept prototype has been implemented to examine the performance of proposed approach using WEKA and MATLAB data mining tools. It is observed that proposed methods gives 96.922% accuracy. A comparative analysis performed between proposed methods and KMSVM (Simple K-mean with SVM classification) and it is observed that proposed method gives better results.

Keywords—IDS; KDD99cup; FS; DBSCAN; K Mean++; SMO.

I. INTRODUCTION

The usage of web augmented altogether the fields**.** Because the usage of web is increasing in our way of life, the network (n/w) security is turning into necessary in sequence to get security, integrity and confidentiality of a resource. Together with the firewalls, IDS has become a main element of the protection system. The position of IDS is to lure the hacker's presence at the network. Because the large range of incidents is growing in our day by day existence IDS's used with improved techniques. IDS plays important role to secure the n/w and its main object is to look at the n/w activities mechanically to identify the malicious attacks. IDS are turning into serious component to secure the n/w in today's world. By exploiting data mining (DM) in IDS will higher the DR, managing the warning False Alarm (FA) rate and reduce FP rate. IDS deals and identifies with n/w resources and malicious n/w of computer. In sequence to detection information target IDS has been categorized as into 2 groups:

A. Host -based IDS(HIDS)
HIDS's planned to observe, response and monitor to movement and attacks on that host.

B. Network-based IDS(NIDS)
NIDS's capture n/w traffic for their IDS operation [1].

II. DATASET DESCRIPTION

KDD'99 cup datasets is utilized for intrusion detection metrics. This data set represented by 41 features. It constitutes 4,940,000 data. Each TCP connection was labeled as "attack" or "normal" with a specific attack kind; the extent of all connection record is 100 bytes. Simulated attack kinds fall in a next four groups[2]:

A. DOS: In this form of attack, the memory useful resource or computing turns into too full or too busy to control denies valid, or valid requests customers get admission to to a device.

B. U2R: Attacker achive to a usual customer account on the computer system and is capable to utilize certain vulnerability to increase root access to the system.

C. R2L: It arises when an attacker who has the ability to transmit packets for machine above a n/w but who doesn't have an account on which machine exploits certain vulnerability to increase local access as a customer of which machine.

D. PROBE Attack: It's effort to collect info about a computers n/w for the apparent purpose of avoiding its security controls. Attacks could be classified depend on the mixture of these 41 features.

III. FEATURE SELECTION

Real time intrusion detection isn't possible as a significance of abundantly of data continuing upon the web. Feature selection (FS) can aid to remove the specific computation besides replica complexity [3]. FS is really a method of selecting a

subset associated with of relevant features thru eliminating many needless and repetitive features from [4] the data for making effective Learning models. Procedure for Feature Selection: FS procedures require four basic stages in a simple FS approach illustrate in diagram 1.

1. Production process in sequence to produce the upcoming applicant subset
2. Estimation function to it can estimate the subset
3. Ending criteria to make a decision when to end.
4. Acceptance process utilized for validates the subset [5].

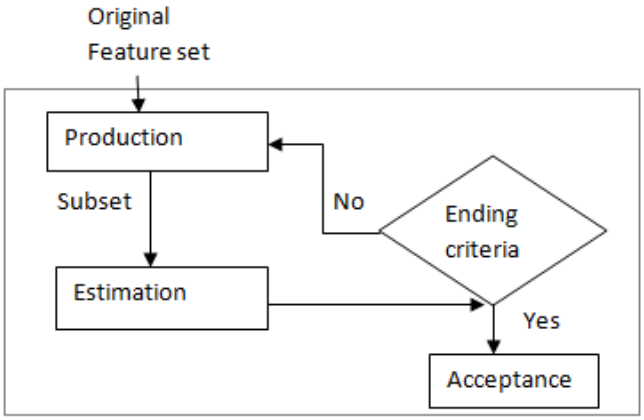

Fig.1Basic FS algorithm

Input (I/P):
KS - Knowledge sample f with attribute F, |F| = x
M - Estimation to be highest
GO – successor production Output:
Result – (weighted) feature set
I: = initial position (F);
Result: = {consistent with M;
Repeat
I: = Search technique (I, GO (M), F);
F':= {consistent with M};
If M (F') =M (resolution) or (M (F') =M (resolution) and |F'| < |resolution|) then
Solution: =F';
Until Stop (M, I).
The filter manner utilizes the selective principles for FS. The correlation statistical or coefficient test as f-test or t-test is explited to filter in the filter FS approach [6].

IV. Density–Based Spatial Clustering of Application with Noise (DBSCAN)

It locates region of maximum density from one another. This algorithm can be utilized to identify clusters of any shape in data containing noise and outliers. Identified clusters will be removed further.
Density = No of points in a particular radius.
Core point = If a point has greater than a particular no. of points within density. Core points are points inside of a cluster.
Border point = This point has less than core points in given density, but it is the nearest of a core point.
Noise point = Noise point is a point which is neither Core point nor Border point
Two or more core points are place within same cluster if they are adjacent within a distance of one another. Any border point that is adjacent enough to a core point is place within cluster as the core point. Noise points are rejected.

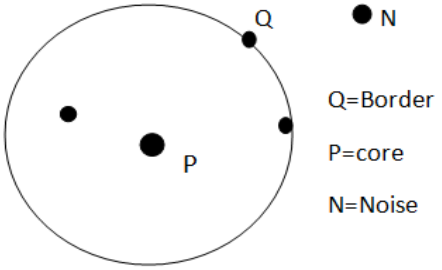

Fig.2 DBSCAN Concept

V. K- Means ++

K-Mean++ is associate algorithmic program for choosing the preliminary values for the K-Mean clustering. K-Mean problematic associate approach of evading the often unhealthy cluster found via the quality K-Mean algorithmic program. Although find a definite result to K-Mean problem for directional I/P unit NP-hard[8], the quality technique to find classified approximate result(often called Lioyd's algorithmic program for K-Mean) is used broadly speaking and regularly discover affordable results.

Following are the two theoretical drawback of K-Mean algorithm are:

- Primary, it's been conferred that the period of time of worst case of the algorithmic rule is super polynomial within the size of I/P.
- Second, the approximation initiate are often every which way unhealthy concerning the target operate equated to the best cluster.

The K-Mean++ algorithmic rule address ensuring obstacles via specifying a method to initialize the cluster center earlier continuing with the standard K-Mean optimization iteration. With the k-means++ data formatting,

the algorithmic rule is assured to observe outcome that's O (log k) competitive to the optimum k-means outcome [7].

VI. SEQUENTIAL MINIMAL OPTIMIZATION (SMO)

The problem of Quadratic programming (QP) difficulty which arrives through SVM training is removed in SMO classification algorithm. SMO is a repeated algorithm for illuminating the optimization obscurity defined above. SMO halts this issue into an order of least feasible sub- issue that are then determination consistently. Owing to the linear fairness constraint comprising the Lagrange multipliers $\alpha 1$, the least feasible problematic includes 2 such multipliers $\alpha 2$ and $\alpha 1$ and. Then, for any 2 multipliers and the constraints are reducing to and this reduces issue is answered analytically: One necessity to discovery a least of a one-dimensional quadratic function.

$$0 \leq \alpha_1, \alpha_2 \leq C,$$

$$y_1\alpha_1 + y_2\alpha_2 = k$$

K is the negative of the summation over the rest of time in the similarity constraint that is fixed in every repetition.

Following are the main steps of SMO algorithm:

a) Locate a Lagrange multiplier that breach the Karush–Kuhn–Tucker α_1 (KKT) conditions for the optimization difficulty.
b) Select a 2nd multiplier α_2 and optimize the pair off $r(\alpha_1, \alpha_2)$.
c) Replicate steps (a) and (b) until meeting.

When everyone Lagrange multipliers convince the KKT situations, the difficulty has been resolved. Though this algorithmic rule is convinced to meet, heuristics are used to elect the couple of multipliers so as to increase the convergence rate. This is risky for huge data sets due to there are of n (n − 1) feasible selections for α_i and α_j.

VII. LITERATURE SURVEY

Saad Mohamed Ali Mohamed Gadal, Rania A. Mokhtar (2017) define a hybrid machine learning approach for NIDS arrangement of K-means clustering followed by SMO classifier. It proposes hybrid technique that capable to decrease the FP alarm rate, FN alarm rate, to better the DR and identify zero-day attackers. The classification has been done by SMO classifier. Next testing and training the define hybrid machine learning method, the outcome have illustrate which define method (K-mean followed by SMO) has attain the FA rate to (1.2%), positive DR of (94.48%) and attain accuracy of (97.3695%)[8].

Angela Denise Landress (2016) The define study will utilize a set of artificial intelligence machine learning approach to decline the measure of FP in ABID records. This technique collects clustering utilizing the K-means algorithm, FS which utilize the J48 Decision Tree algorithm, and self-establishing maps to efficiently diminish FP used by KDD CUP 99 data set [9].

Hatim Mohamad Tahir et. al (2016) The intention of this research is to enhanced the existing ABID technique utilizing K-Means clustering process as to increase the DR and accuracy rate whereas decreasing FA. The trouble with outliers may interrupt the K-Means clustering procedure as it might be avoided in the clustering process from mixing with the normal data that make the NIDSs become less accurate. Thus this research aims to better the process of the ABID systems that balance the loss of information or ignored data in clustering. An integrated machine learning technique using K-Means Clustering with discretization method and Naive Bayes Classifier (KMC-D+NBC) is proposed against ISCX 2012 assessment of Intrusion Detection Dataset. The result depicts that the define approach method generates better DR i.e. 99.3% and accuracy i.e. 99.5% and decrease the *FA* to 1.2% with enhanced efficiency of 0.03 seconds time utilize to make this approach [10].

V. Jyothsna , V.V. Rama Prasad (2016) proposed canonical correlation analysis to optimize the features towards detecting the intrusions. The optimal election features simplifies the procedure of FAIS which is utilized in our earlier work article. The research was whole exploiting a data set of benchmark. The outcomes establish and promising to elect optimal attributes of the n/w transactions utilized for training. Moreover the define replica diminished the process complexity and completion time and retains the maximal prediction accuracy [11].

Nutan Farah Haq et. al (2015) In this paper an ensemble IDS system is define via a sequence of machine learning classifiers and a hybrid FS approach. This build a appropriate NSL-KDD train dataset, decrease features into 12 from 41 through the describe hybrid FS method; Building up classification models (Naive Bayes, Bayesian n/w and J48) using training data; Classification of test instances using majority vote and earlier built classification models as base classifier. The FP)rate of the define replica is 0.021 with a TP rate of 98.0%.The outcome demonstrates that the proposed ensemble model is a consistent and exceed other classifiers performance [12].

VIII. PROPOSED WORK

A novel approach has been proposed to increase the classification accuracy of KDD dataset. Proposed approach consists of following four phases shown in the figure 3.

1. Noise Reduction
2. Feature selection
3. Clustering
4. Classification

Noise Reduction: KDDCUP 99 intrusion detection dataset suffered by redundant data and noise, therefore it is important to remove noise before further processing. DBSCAN method is utilized for noise removal. Here first 1 lakh record of KDD

cup dataset has been supplied to DBSCAN methods for noise removal.141 records are identified as noise by BDSCAN algorithm.

Feature selection: KDDCUP 99 dataset consists of 41 attribute. In which some are relevant and some are irrelevant therefore it is required to remove irrelevant attribute in order to improve classification accuracy. Genetic search technique has been utilized for feature selection work. Here out of 41 features only 13 relevant attributes are selected for further processing, shown in the table 1.

Clustering: In this phase K-Mean ++ algorithm has been applied on data set. K-mean++ generate five clusters as an outcome, shown in table 2.

Classification: In this phase SMO (Sequential minimal optimization) classification algorithm applied on clustered data which classifies whether intrusion present or not, shown in table 3.

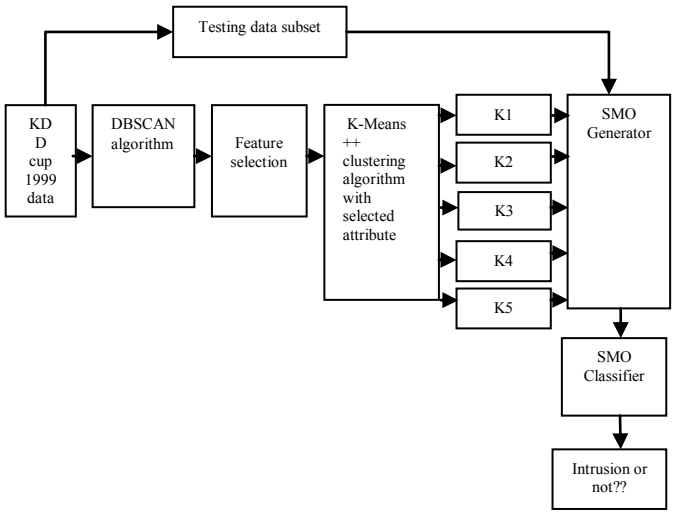

Fig. 3 Block Diagram of Proposed Architecture

IX. Experimental Results and Analysis

Experimentation has been carried out using WEKA 3.6.15 and MATLAB R2013b (8.2.0701) data mining tool.

Following table 1 shows the description of selected attribute from the whole attributes in dataset.

Table 1: selected features

Attribute No	Attribute Name
2	protocol_type
3	service
5	src_bytes
6	dst_bytes
7	Land
9	Urgent
12	Logged in
14	lroot_shell
23	Count
30	diff_srv_rate
31	srv_diff_host_rate
32	dst_host_count
37	dst_host_srv_diff_host_rate

Table 2: K-MEAN ++ outcomes

S.NO	NAME	No of record
1	K1	79233
2	K2	830
4	K3	8
3	K4	230
5	k5	80301

Table 3: SMO outcomes on clustered data

Data	Correct categorized instances	Incorrect categorized instances	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error	Total Number of Instances
K1	99.918 %	0.082 %	0.9984	0.2224	0.2724	66.825%	66.8012%	79232
K2	96.988 %	3.012 %	0.9322	0.2289	0.2842	75.876%	73.23 %	830
K3	100%	0%	1	0	0	0%	0%	8
K4	87.826 %	12.1739%	0.7953	0.2075	0.3066	118.449%	104.258%	230
K5	99.882 %	0.1171 %	0.9977	0.2225	0.2726	66.777%	66.8002%	80300
Average	96.922 %	3.0768 %	0.94	0.3313	0.2271	65.5742%	62.2178%	32120

X Comparative Analysis

A comparative analysis has been carried out between proposed approach and KMSVM. Here KMSVM is an approach in which features selection, clustering using k-means and SVM classified are utilized for classification. It is observed that (Hybrid approach of feature selection, DBSCAN, K-Mean++clustering and SMO classification) proposed methods generate better result than KMSVM.

Table 4: Accuracy Outcomes for Reduced Attribute Set:

S.No	Dataset	KMSVM	Proposed
1	K1	89.44%	99.918 %
2	K2	97.69%	96.988%
3	K3	100.00%	100.00%
4	K4	67.21%	87.826%
5	K5	65.91%	99.882%
Average		84.062%	96.922%

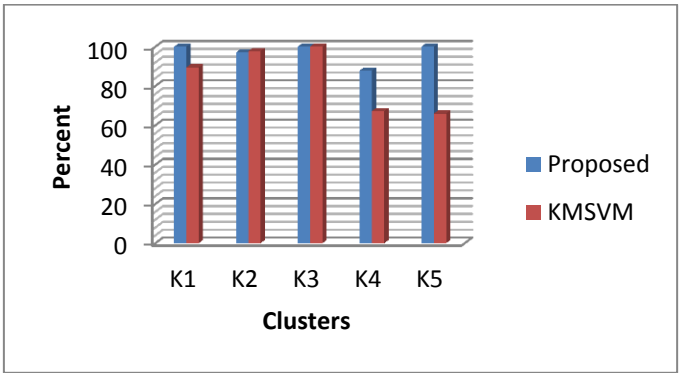

Fig. 4 Accuracy Graph

Fig. 4 is the comparison of Proposed with KMSVM algorithm in terms of accuracy.

XI. CONCLUSION

Security is the most concerning issue in the modern time. In this report a novel approach is proposed which is a hybrid model of classification and clustering.KDD'99Cup dataset suffered from noise which is removed by using DBSCAN algorithm than feature selection technique applied to select relevant attribute. Clustering algorithm applied to cluster data set into K1, K2, K3, K4, and K5. At last classification algorithm applied on clustered data to know intrusion happened or not. The overall approach improves the accuracy. It is observed that obtained accuracy is 96.922% which is better than KMSVM.

References

[1] Vinutha H.P, Dr.Poornima B "A Survey - Comparative Study on Intrusion Detection System" International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 7, July 2015.

[2] 2. M. Tavallaee, E. Bagheri, Wei Lu, and Ali A. Ghorbani, " A Detailed Analysis of the KDD CUP 99 Data Set " , Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defence Applications (CISDA 2009)

[3] Lewis, P. M., "The characteristic selection problem in recognition system", IRE Transaction on Information Theory, 8, PP 171-178, IEEE 2003

[4] George H John, Ron Kohavi and Karl PEger, "Irrelevant Features and the Subset Selection Problem", Proc. of the 11th International Conference. On Machine Learning, Morgan Kaufmann Publishers, pp 121-129, 1994

[5] Harvinder Singh, Prof. Dharminder Kumar "A study on Performance analysis of various feature selection techniques in intrusion detection system" International Journal of Advance Research in Computer Science and Management Studies, Volume 3, Issue 6, June 2015.

[6] S.Vanaja, K.Ramesh kumar "Analysis of Feature Selection Algorithms on Classification: A Survey" International Journal of Computer Applications (0975 – 8887) Volume 96– No.17, June 2014..

[7] Noticewala Maitry, Dinesh Vaghela "Survey on Different Density Based Algorithms on Spatial Dataset" International Journal of Advance Research in Computer Science and Management Studies, Volume 2, Issue 2, February 2014.

[8] Saad Mohamed Ali Mohamed Gadal, Rania A. Mokhtar, "Anomaly Detection Approach using Hybrid Algorithm of Data Mining Technique", 2017 International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE), Khartoum, Sudan

[9] Angela Denise Landress, "A Hybrid Approach to Reducing the False Positive Rate in Unsupervised Machine Learning Intrusion Detection", SoutheastCon, 2016, IEEE

[10] Hatim Mohamad Tahir et. al , "Improving k-means clustering using discretization technique in network intrusion detection system",2016 3rd International Conference On Computer And Information Sciences (ICCOINS)

[11] V. Jyothsna, V.V. Rama Prasad, "FCAAIS: Anomaly based network intrusion detection through feature correlation analysis and association impact scale",The Korean Institute of Communications Information Sciences, ICT Express 2 (2016) 103–116

[12] Nutan Farah Haq et. al, "An Ensemble Framework of Anomaly Detection using Hybridized Feature Selection Approach (HFSA)", SAI Intelligent Systems Conference 2015 November 10-11, 2015 | London, UK.