*Tariq et al LGURJCSIT 2019*

LGU International Journal for
**Electronic Crime Investigation**

LGU (IJECI)
ISSN: 2522-3429 (Print)
ISSN: 2616-6003 (Online)

Research Article Vol. 3 Issue 3, July - September 2019

# Anomaly based Intrusion Detection System

**Muhammad Arslan Tariq[1], Rehmatullah[2], Waheed-ul-Hassan[3]**
arslan.tariq@lgu.edu.pk[1], rehmatullah@lgu.edu.pk[2], waheedhassan@lgu.edu.pk[3] Lahore Garrison University

**Abstract:**
In the digital World full of hackers and scammers, data security is what everyone needs the most. Hackers and scammers invent new ways of stealing information on daily basis. A method to come up with more precise system is Intrusion Detection system. IDS is todays need because, it helps the individuals to keep up their confidentiality and integrity. Intrusions, that disturbs the security and secrecy of the system, has become major concern for many organizations. The logic and ways Intrusion Detection System uses are related to these days. Through cloud computing, Intrusion Detection System has creäted a world where it can flourish and be most operative. By means of cloud computing, the fundament has engrossed with the Intrusion Detection technology.

**Keywords**— Anomaly Based IDS, IDS, Poly kernel, Normalized poly kernel RBF kernel

## I. Introduction

An Intrusion Detection System is employed to differentiate every kind of vulnerable links available in traffic and systems that cannot be identified with traditional firewall. It includes network attacks against susceptible facilities, attacks on data driven applications and host-based networks such as enhancement in privilege, unlicensed access and logins to sensitive data and vulnerable files (i.e. Malwares, Trojans, and worms).

### A. Need for Intrusion Detection System

Confidentiality is the main concern between individual and corporate sector. If these problems are not solved, businesses will not be in a better position to truly take advantage of all.

Like, Sony Pictures Entertainment [1] experienced one of the most vulnerable commercial attack in the history. Thousands of records grabbed by hackers were revealed online with personal details of almost 6,000 employees of Sony, including Sony feature films and the salary details of top management. The hackers also achieved to retrieve details about Deloitte financiers who are Sony's auditors.

The above-mentioned data breach, that happened on 24th November, caused in the halt of whole computer network of one of Hollywood's prime and most authoritative studios. Here is the collective report that the hacking was carried out by North Korea in payback for the future release of a Sony comedy movie called "The Interview". The storyline tracks Seth Rogan and James Franco who were working in the CIA (Central Intelligence Agency) to eliminate Kim Jong-un the dictator of North Korea.

Industrial think tanks observed that almost 60,000 new, vulnerable computer programs and 315,000 new, vulnerable files are discovered daily. From 2006 to 2012, the number of security happenings stated by federal agencies amplified from 5,503 to 48,562 - a rise of 78.2% - and in 2013 McAfee investigation estimated that worldwide cybercrime failures might total $400 billion. Cyber-attacks are a risk to America's nationwide and financial security, in addition to separate privacy, to the fundamental and most important factor, corporate strategies, and

knowledgeable property for all [8].

Cloud computing, though, has taken new applicability to IDS structures, resulting flow in the IDS marketplace. An important element of today's security top preparations, Intrusion Detection Systems are created to sense attacks that can happen regardless of preventive procedures. In fact, Intrusion Detection System is today's unique top selling security equipment and it is predicted to remain increase. Despite everything, cloud security is far too multifaceted to be checked physically.

This study deals with anomaly-based intrusion detection system. It uses support vector machine for model evaluation.

### B. Support Vector Machine

Support vector machine (SVM) is the best-recognized algorithm for classification of binary data. It uses statistical learning method for classification and regression by using different kernel functions. Its applications include a wide range of pattern recognition applications and now it is popular in networks security due to good generality nature and to overcome the curse of dimensionality. The SVM selected the appropriate parameters for model evaluation.

### C. Limitations of SVM

SVM is a supervised learning model required labelled data for learning. It is designed for Binary classification [14]. Another issue is training of support vector machine (SVM) is a time-consuming process and required a huge dataset. Thus, it is computationally costly, and resource restricted for informal networks, that increase the architecture complexity and decrease accuracy [10].

To resolve this issue NSL-KDD binary dataset is used where data is labelled as normal or Anomaly only.

## II. Literature Review

Computer world is growing explosively. Computer System suffer security vulnerabilities that are technically difficult and economically costly. On KDD, test set is a classification rate of 86% to nearly 100%.

There square measures some issues within the KDD knowledge set that cause the analysis results on this knowledge set to be dishonest. That square measure mentioned below:

One of the foremost vital insufficiencies within the knowledge discovery in database's (KDD) dataset is that the immense variety of redundant stored information, that causes the training algorithms to be projected towards the frequent records, So, to stop them from learning uncommon records that square measures typically additional harm to networks like R2L and U2R attacks. Furthermore, the presence of these frequent records within the check set can cause the assessment results to be biased by the strategies that have higher detection rate on the repeated records [3].

Solution for this is to first take away all the redundant records in each training and testing set. Moreover, to make a different set of the knowledge discovery in databases (KDD) knowledge set, we have a tendency to willy-nilly sampled records from the #successfulPrediction price teams, in such how that the numeral of records chosen from every cluster is reciprocally proportional to the proportion of records within the original #successfulPrediction price teams. for example, the quantity of records within the #successfulPrediction price cluster of the KDD toy constitutes zero.04% of the initial records, therefore, 99.96% of the records during this cluster square measures enclosed within the generated sample. The generated knowledge sets, square measure KDDTrain+ and KDDTest+.

Dataset social control is important to boost the performance of IDS once amount of dataset is large. Hence, technique used is Min-Max technique of social control.

Features will be selected based on information gain. It was calculated as

Let [5] S be a group of training set samples with their match up labels. Imagine there are m categories/classes and the training set contains si samples of category/class I and s are that the total variety of samples within the training set. predictable data required to classify a sample is computed by:

Let Sj contain sij samples of class/category i. A feature F with values will divide the training set into v subsets wherever Sj is that the set that has the worth fj for feature F [5]. moreover, Entropy of the feature F is calculated as

$$I(s_1, s_2, s_3, \ldots\ldots, s_m) = -\sum_{i=1}^{m} \frac{s_i}{s} \log_2 \left(\frac{s_i}{s}\right)$$

Information gained for F is calculated as:

$$E(F) = \sum_{j=1}^{v} \frac{s_{1j} + \cdots + s_{m_j}}{s} * I(s_{1_j}, \ldots\ldots, s_{m_j})$$

The dependency magnitude relation [6] is solely calculated therefore Dependency ratio

$$Dependency\ Ratio = \frac{HVF}{TIN} - \frac{OTH}{TON}$$

Where
HVF = highest variety of occurrence variation for a category label in attribute f.
TIN = total variety of occurrences of that category within the dataset
OTH = variety of occurrences for different category labels supported a or a group of Variations.
TON = total variety of instances of category/class labels within the dataset creating OTH.

It helps to pick out options by high worth to low worth and so they're evaluated.

Rule induction is [15] one in all the major varieties of data processing and in unsupervised learning systems it is probably the most common variety of information discovery. Rule induction on a data is a vast responsibility wherever all doable patterns are completely force out of the information.

For the how much the rule to be helpful there must be two things that provide a great information

- o Accuracy - however typically is that the rule corrects?
- o Coverage - however typically will the rule apply?

## III. NSL-kdd dataset

The dataset employed in the study is NSL-KDD. NSL-KDD could be a dataset counseled to resolve a number of the characteristic issues of the KDD'99 dataset [9]. Although, this new sort of KDD information set still suffers from a number of the issues mentioned by McHugh and won't be an ideal demonstration of current real networks, attributable to the deficiency of public datasets for network-based Intrusion Detection Systems, it is still sensible as a good benchmark dataset to assist researchers to compare completely different intrusion detection ways.

Moreover, the quantity of records within the NSL-KDD train and test set are affordable. This advantage makes it cheap to run the experiments on the whole set while not the requirement to arbitrarily choose a little portion. Therefore, analysis results of various analysis work are reliable and comparable.

| Sets | No of records |
|---|---|
| NSL-KDDTest+ | 22544 |
| NSL-KDDtrain+ | 125973 |

Features of datasets are:

| Sr. | Features Name |
|---|---|
| 1 | Duration |
| 2 | Protocol type |
| 3 | Service |
| 4 | Flag |
| 5 | src_bytes |
| 6 | Des bytes |
| 7 | Land |
| 8 | Wrong fragment |
| 9 | Urgent |
| 10 | Hot |
| 11 | Failed logins |
| 12 | logged in |
| 13 | Num compromised |
| 14 | root shell |
| 15 | Su attempted |
| 16 | num root |
| 17 | num file_creations |
| 18 | num shells |
| 19 | num access files |
| 20 | num_outbound_cmds |
| 21 | is_host_login |
| 22 | is_guest_login |
| 23 | Count |
| 24 | Srv count |
| 25 | Serror rate |
| 26 | Srv serror_rate |
| 27 | Rerror rate |
| 28 | Srv rerror_rate |
| 29 | Same srv_rate |
| 30 | Dif srv_rate |
| 31 | Srv diff_host_rate |
| 32 | Dst host_count |
| 33 | Dst host_srv_count |
| 34 | Dst host_same_srv_rate |
| 35 | Dst host_diff_srv_rate |
| 36 | Dst host_same_src_port_rate |
| 37 | Dst host_srv_diff_host_rate |
| 38 | Dst host_serror_rate |
| 39 | Dst host_srv_serror_rate |
| 40 | Dst host_rerror_rate |
| 41 | Dst host_srv_rerror_rate |
| 42 | Label |

Table 1- Features of Dataset

## IV. Working of Anomaly Based Intrusion Detection System:

Anomaly based refer to the statistical measure of system features. For this NSL-KDD dataset is used.

In general, Anomaly based detection involves following steps:

### 1. Pre-Processing

It is an important step in data mining process. It converts the raw into understandable format. There it required a training dataset for the learning of IDS. It contains 41 features.

### 2. Feature Selection

In machine learning it is a procedure of choosing a subset of pertinent features/attributes used to create model. For specific results we need relevant features. Feature selection methods are adopted for following motives

- Over simplification of model so it becomes easy to understand.
- Shorter training time.
- To avoid curse of dimensionality.
- Enhance generalization by reducing over fitting.

Feature are selected using Cfs Subset Evaluation with Genetic Search algorithms. Attributes are selected using percentage folds.

number of folds (%) attribute

| No of fold percentage | Attributes Name | Sr# of features |
|---|---|---|
| 1(10 %) | duration | 1. |
| 0( 0 %) | protocol_type | 2. |
| 5( 50 %) | service | 3. |
| 8( 80 %) | flag | 4. |
| 10(100 %) | src_bytes | 5. |
| 10(100 %) | dst_bytes | 6. |
| 3( 30 %) | land | 7. |
| 2( 20 %) | wrong_fragment | 8. |
| 0( 0 %) | urgent | 9. |
| 0( 0 %) | hot | 10. |
| 0( 0 %) | num_failed_logins | 11. |
| 10(100 %) | logged_in | 12. |
| 1( 10 %) | num_compromised | 13. |
| 1( 10 %) | root_shell | 14. |
| 0( 0 %) | su_attempted | 15. |
| 0( 0 %) | num_root | 16. |
| 0( 0 %) | num_file_creations | 17. |
| 3( 30 %) | num_shells | 18. |
| 3( 30 %) | num_access_files | 19. |
| 0( 0 %) | num_outbound_cmds | 20. |
| 0( 0 %) | is_host_login | 21. |
| 3( 30 %) | is_guest_login | 22. |
| 3( 30 %) | count | 23. |
| 0( 0 %) | Srv count | 24. |
| 7( 70 %) | Serror rate | 25. |
| 7( 70 %) | Srv serror_rate | 26. |
| 2( 20 %) | Rerror rate | 27. |
| 1( 10 %) | Srv rerror_rate | 28. |
| 10(100 %) | Same srv_rate | 29. |
| 7( 70 %) | Diff srv_rate | 30. |
| 8( 80 %) | Srv diff_host_rate | 31. |
| 4( 40 %) | Dst host_count | 32. |
| 3( 30 %) | Dst host_srv_count | 33. |
| 4( 40 %) | Dst host_same_srv_rate | 34. |
| 0( 0 %) | Dst host_diff_srv_rate | 35. |
| 2( 20 %) | Dst host_same_src_port_rate | 36. |
| 4( 40 %) | Dst host_srv_diff_host_rate | 37. |
| 5( 50 %) | Dst host_serror_rate | 38. |
| 7( 70 %) | Dst host_srv_serror_rate | 39. |
| 0( 0 %) | Dst host_rerror_rate | 40. |
| 2( 20 %) | Dst host_srv_rerror_rate | 41. |

**Table 2-** Selected Features of Dataset

Features selected are:

| No of fold percentage | Features selected | No of features |
|---|---|---|
| 100 | 5,6,12,29 | 4 |
| >50 | 3,4,5,6,12,25,29,30,31,34,38,39 | 12 |
| >0 | 1,3,4,5,6,7,8,12,13,14,18,19,22,23,24, 25,26,27,28,29,30,31,32,33,34,36,37, 38,39,41 | 29 |

**Table 3.** Feature selected on the basis of fold percentage


**Graph 1-** Features Selected

### 3. Parameter Optimization

It is a process of choosing optimal parameter for learning algorithms. This measure is known as hyperparameter and resultant model solves problem optimally.

### 4. Classification

Classification is a process of arrangement of optimized parameters so that useful information can extract in data. It assigns items in a

collection to categories or classes. It results in the formation of a model.

In machine learning, modelling SVM's are supervised learning models with linked learning algorithms that examine facts used for classification or regression study.

Models are developed using SMO. Sequential minimal optimization (SMO) is a process for elucidation the quadratic programming (QP) problem that rises during the learning of support vector machines. Following kernels are to be used.

**Poly kernel** is said to be polynomial kernel. It finds the similarities not only between features but also among there subsets. Polynomial kernel is defined as:

$$K(x, y) = (x^t y + c)^d$$

In this,
x , y = vectors in vector space
c= effect of higher degree order term vs lower degree term. C always greater than 0. If c=0 then kernel is said to be homogenous.

**Normalized poly kernel** is the refined form of Polynomial kernel. First data is normalized and then processed. It is defined as:

## 5. Evaluation

Model will be evaluated on the bases of confusion matrix. Multiple scores are measured such as: accuracy, precision, recall, F-measure by performance of 10-fold cross-validation.

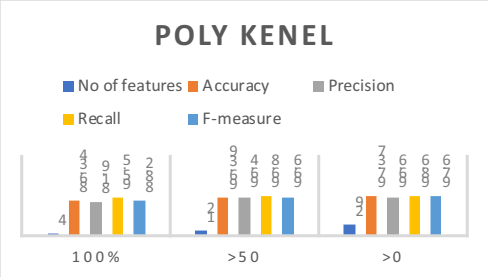
## V. Result

This proposed study of IDS is tested using WEKA (Waikato Environment for knowledge Analysis).
The dataset NSL-KDD has advantages over KDD99 due to Removal of redundant records and affordability for use in experimental purpose. Classification results are based on NSL_KDD 20%. The cross-validation folds are set to 10.

For Poly kernel

| No of fold percentage | No of features | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|---|
| 100 % | 4 | 0.8634 | 0.819 | 0.955 | 0.882 |
| >50 | 12 | 0.9639 | 0.964 | 0.968 | 0.966 |
| >0 | 29 | 0.9737 | 0.966 | 0.986 | 0.976 |

**Table 4-** Poly Kenel

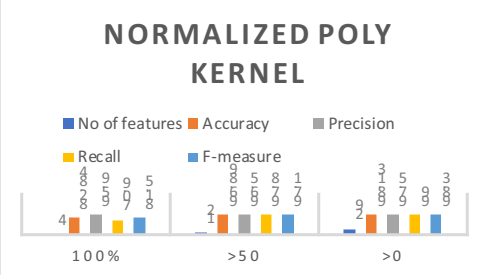



**Graph 2.** Poly kernel Evaluation results.

For Normalized Kernel

| No of fold percentage | No of features | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|---|
| 100 % | 4 | 0.8284 | 0.959 | 0.709 | 0.815 |
| >50 | 12 | 0.9689 | 0.965 | 0.978 | 0.971 |
| >0 | 29 | 0.9813 | 0.975 | 0.99 | 0.983 |

**Table 5-** Normalized Kernel

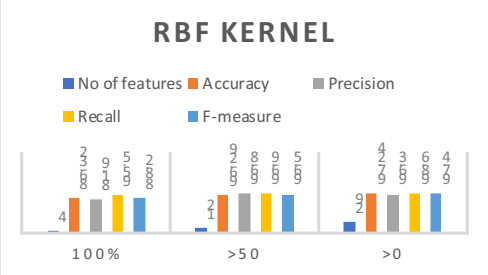



**Graph 3-** Poly Kernel

For RBF kernel

| No of fold percentage | No of features | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|---|
| 100 % | 4 | 0. 8632 | 0.819 | 0.955 | 0.882 |
| >50 | 12 | 0.9629 | 0.968 | 0.969 | 0.965 |
| >0 | 29 | 0.9724 | 0.963 | 0.986 | 0.974 |

**Table 6-** RBF kernel

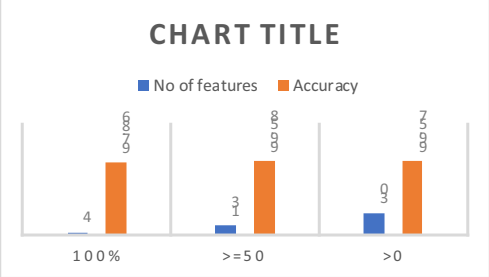



**Graph 4-** RBF Kernel

For Decision tree (J48)

| No of fold percentage | No of features | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|---|
| 100 % | 4 | 0.9786 | 0.965 | 0.997 | 0.98 |
| >=50 | 13 | 0.9958 | 0.995 | 0.997 | 0.996 |
| >0 | 30 | 0.9957 | 0.996 | 0.996 | 0.996 |

**Table 7-** Decision Tree

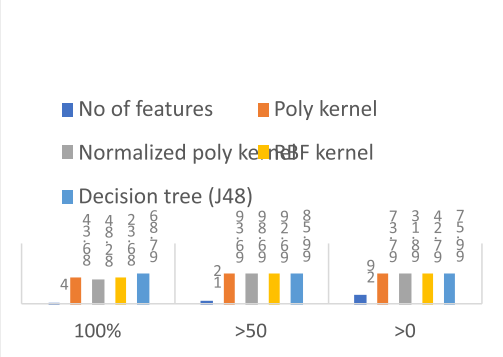



**Graph 5-** Decision Accuracy

As the no of attributes increases the accuracy increases to some extent. The normalized poly kernel achieved high accuracy then other SMO kernels. This classification evaluation is binary class evaluation.

| No of fold percentage | No of features | Poly kernel | Normalized poly kernel | RBF kernel | Decision tree (J48) |
|---|---|---|---|---|---|
| 100 % | 4 | 86.34 | 82.84 | 86.32 | 97.86 |
| >50 | 12 | 96.39 | 96.89 | 96.29 | 99.58 |
| >0 | 29 | 97.37 | 98.13 | 97.24 | 99.57 |

**Table 8-** Binary class Evaluation




**Graph 6-** Comparison

## VI. Conclusion

IDS is today's need because, it helps the individuals to keep up their confidentiality and integrity. Intrusion that disturbs the security and secrecy of the system, has become major concern for many organizations.

Hence, there's a desire of strong IDS which might observe completely different attack with high attack recognition accuracy. In this, we've got proposed a technique of intrusion detection using SVM which might increase the intrusion detection correctness.

## VII. References


[1] Newsweek. (2017). Sony Cyber Attack One of Worst in Corporate History. [online] Available at:

[2] http://www.newsweek.com/sony-cyber-attack-worst-corporate-history-thousands-files-areleaked-289230 [Accessed 22 Dec. 2017].

[3] A. A. Rao "A Java Based Network Intrusion Detection System (IDS)".

[4] M. Tavallaee, E. Bagheri, W. Lu, A. A. Gorbani "A detailed analysis of KDD CUP 99 dataset"

[5] J. McHugh, "Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory," ACM Transactions on Information and System Security, vol. 3, no. 4, pp. 262-294, 2000.

[6] H. G. Kayac?k, A. N. Z. Heywood, M. I. Heywood "Selecting Features for Intrusion Detection: A

[7] Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets"

[8] A. A. Olusola., A. S. Oladele. and D. O. Abosede "Analysis of KDD '99 Intrusion Detection Dataset for Selection of Relevance Features" Proceedings of the World Congress on Engineering and

[9] Computer Science 2010 Vol I WCECS 2010, October 20-22, 2010, San Francisco, USA ISBN: 978988-17012-0-6

[10] Byunghae, C., kyung, W.P. and Jaittyun, S. (2005) Neural Networks Techniques for Host Anomaly Intrusion Detection using Fixed Pattern Transformation in ICCSA. LNCS 3481. 254-263.

[11] Anon, (2017). Saggi e Memorie di storia dell'arte. [online] Available at: http://www.ccianet.org/wpcontent/uploads/2014/04/Cybersecurity.pdf [Accessed 22 Dec. 2017]

[12] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," Submitted to Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2009.

[13] J.F Joseph,A. Das,B.C. Seet, (2011) Cross-Layer Detection of Sinking Behavior in Wireless Ad Hoc Networks Using SVM and FDA. IEEE Transaction on dependable and securecomputing, Vol. 8, No. 2, MarhApril 2011.

[14] T. Shon, Y. Kim, C. Lee and J. Moon, (2005), A Machine Learning Framework for Network Anomaly Detection using SVM and GA, Proceedings of the 2005 IEEE.

[15] S. Peddabachigari, A. Abraham, C. Grosan, J. Thomas (2005). "Modeling Intrusion Detection Systems using Hybrid Intelligent Systems." Journal of Network and Computer Applications.

[16] R.C. Chen, K.F Cheng and C. F Hsieh (2009),using support vector machine and rough set for network intrusion system.

[17] K. T. Khaing (2010),"Recursive Feature Elimination (RFE) and k-Nearest Neighbor (KNN) in SVM."

[18] J. Han and M. Kamber, "Data Mining Concepts and Techniques" Morgan Kaufmann publishers .an imprint of Elsevier, ISBN 978-1-55860-901-3. Indian reprint ISBN 978-81-312-0535-8.

[19] G. P. Dubey, Prof. N. Gupta, R. K. Bhujade "A Novel Approach to Intrusion Detection System using Rough Set Theory and Incremental SVM". International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume 1, Issue 1, March 2011.

[20] A. A.Olusola, A. S. Oladele and D. O. Abosede, "Analysis of KDD '99 Intrusion Detection Dataset for Selection of Relevance Features. Proceedings of the World Congress on Engineering and Computer Science 2010 Vol. I WCECS 2010, October 20-22, 2010, San Francisco, USA.

[21] B. Hur, Asa, Horn, David, Siegelmann, Hava, and Vapnik, Vladimir; "Support vector clustering" (2001) Journal of Machine Learning Research, 2: 125-137.