Network Intrusion Detection: Comparative Analysis of NSL-KDD and CIC-IDS2017 Datasets

Ritesh Kumar¹, Gagandeep Jagdev²*

¹Research Scholar, M.Tech (CE), Yadavindra College of Engineering, Talwandi Sabo. Punjab, India.
²Technical Officer, Punjabi University Guru Kashi Campus, Damdama Sahib, Talwandi Sabo, Punjab, India.

*Corresponding Author: Gagandeep Jagdev, Technical Officer, Punjabi University Guru Kashi Campus, Damdama Sahib, Talwandi Sabo, Punjab, India

Abstract: The research paper explores NSL-KDD and CIC-IDS2017 datasets and implementing multiple encoding techniques on these datasets. The three categories of encoding techniques implemented on NSL-KDD dataset are One Hot Encoding, Removal of missing values, and Removal of Categorical Data. The six categories of encoding techniques implemented on CIC-IDS2017 are Eliminating categorical data or missing values, Imputing mode for missing values, Imputing median for missing values, Imputing mean for missing values, Imputing 0 for missing values, and Simple imputer. Thereafter the true labels (Binary True labels and Multiclass True labels) are implemented. The clustering is performed on the datasets using K-Means, Isolation Forest, DBSCAN, and Local Outlier Factor. For NSL-KDD, the Score Metrics Menu with three options to calculate among F1 score, Normalized Mutual Info Score, and Adjusted Rand Score using the appropriate Average Method is implemented and the score of the desired parameter is evaluated. For CIC-IDS2017, the Score Metrics Menu with three options to calculate among F1 score, Normalized Mutual Info Score, and Adjusted Rand Score using the appropriate Average Method is encountered and the score of the desired parameter is evaluated.

Keywords: CIC-IDS2017, F1-score, Intrusion Detection System, K-Means, NSL-KDD.

1. INTRODUCTION

An Intrusion Detection System (IDS) is a framework that screens network traffic for dubious action. IDS issues an alert once it finds a doubtful entry in the network. A software application filters an organization or a framework for malicious activity or strategy breaking (Abbas et al., 2022). The malicious activity is reported to the organization for further action as it is collected using a framework. A SIEM (Security Information and Event Management) framework coordinates yields from numerous sources and uses caution sifting procedures to separate pernicious movement from deceptions. Even though interruption recognition frameworks screen networks for possibly malevolent action, they are additionally arranged for phony problems. Thus, associations need to tweak their IDS items when they initially introduce them (Devulapall, 2021). It implies appropriately setting up the interruption identification frameworks to perceive what ordinary traffic in the organization resembles when contrasted with noxious action. Interruption anticipation frameworks additionally screen network parcels inbound the framework to check the noxious exercises engaged with it and without a moment's delay send the admonition warnings.

2. MACHINE LEARNING CLASSIFIERS

The research work focuses on four prominent machine learning classifiers briefly explained as under.

2.1. K-Means

K-Means compute the distance between points and group centers. It requires fewer computations and has linear complexity of $O(n)$. As K-Means make use of centroids, it assumes the clusters to be circular-shaped. K-means clusters data by beginning with user-specified K initial cluster centroids. Depending on the assignments, the algorithm recalculates the centers of the clusters and reassigns the points to the nearest cluster center. When the process terminates, the data gets clustered into K clusters using K-Means (Jadhav & Pellakuri, 2021). The sum of the squared distances to cluster centers with different values of K is conducted (Ravikumar, 2021).
2.2. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Unlike K-Means, DBSCAN considers the density of data into account for clustering. The execution time of DBSCAN depends on the size of the dataset and the configuration and resources of the machine. The DBSCAN is comparatively slower than K-Means as it groups together the points that are near to one another based on distance measurement.

2.3. Isolation Forest

Isolation forest returns a score for every point representing the possibility of the specific point for being an anomaly and the user is supposed to choose a threshold for which scores will indicate a normal instance and for which an anomaly. The concept of an Isolation Forest is that anomalies are not that frequent and distinguish easily from normal points and hence need lesser partitions for isolation as compared to normal points. The average number of splits decides the score grounded on the number of splits. A score close to 1 is considered an anomaly and a score close to 0.5 is considered as normal point.

2.4. LOF (Local Outlier Factors)

LOF initiates with a parameter K or any arbitrary integer chosen by the user. The algorithm is supposed to evaluate the reach distance to every point, which is the distance from a particular point to another point with a minor warning for close points. Thereafter, the average of the reach distances for a specific point to each of that point’s k-nearest neighbors is calculated.

3. RESEARCH METHODOLOGY

This section elaborates on the designed framework and the methodology adopted for conducting the research work. Fig. 1 shows the flowchart depicting the adopted methodology for conducting the research.

![Flowchart depicting the proposed research methodology](Image)
4. IMPLEMENTATION AND RESULTS

This section depicts the conducted implementation of the proposed methodology. The implementation has been done for executing the anticipated technique on two popular IDS datasets: NSL –KDD and CICIDS 2017.

Fig 2 shows the initiating stage of the conducted work. The option is asked by the user to choose from NSL-KDD or CICIDS 2017.

Fig 2. Initiating stage of the conducted work

Fig. 3 shows option 1 as the selected option. Once the option is selected, the developed framework asks for the path of the dataset file.

Fig 3. Developed program asking for options and path of the dataset file

Fig. 4 shows the uploaded dataset file. After uploading the dataset, the user is asked whether the dataset comprises feature names. Once done with this, the dataset is read and the user is asked to opt for one option from the three available options. The first option is to analyze the dataset with categorical data using Onehot encoding. The second option is for analyzing the dataset after removing the categorical data. The third option is to analyze the dataset with the risk values replacing the server type and flag features.
Fig 4. Menu to select the option for reading the dataset

Fig. 5 shows the selected option from the available menu.

Fig. 6 shows the displayed Encoding Menu for the selection of the labels after entering “1” as a selected option from the Variables Menu. Encoding Menu displays two options. The first option is for “Binary True Labels” and the second option is for “Multiclass True Labels”. The “Binary True Labels” is further categorized into “normal” and “abnormal”. The “Multiclass true Labels” is categorized into “normal”, “DoS”, “Probe”, “R2L”, and “U2R”.

Path of the file: C:\Users\pavilion\.spyder-py3\ritesh\Dataset\KDDTrain+.csv

Dataset has feature names[y/n]: y

Reading Dataset...

*******************************************************
Variables Menu
*******************************************************
1. Data set with categorical data oneHot encoded
2. Data set with categorical data removed
3. Data set with Risk Values replacing Server Type and Flag Features; Protocol Data oneHot encoded

Enter option :

Fig5. Selected option from the available menu
Network Intrusion Detection: Comparative Analysis of NSL-KDD and CIC-IDS2017 Datasets

Fig. 6. Figure depicts the “Encoding Menu” generated post “Variables Menu”

Fig. 7 shows that option 1 has been selected from the “Encoding Menu”. The data is successfully One Hot Encoded. After encoding, the user is asked whether to scale the data or not.

After scaling the data, the user is asked whether to shuffle the data or not as shown in Fig. 8.
Enter option : 1

Data has been successfully One Hot Encoded

Scale data [y/n]: y

Data has been successfully scaled.

Shuffle data [y]/[n]:

---

**Fig 8.** Figure depicts the user being asked whether to shuffle the data or not

Fig. 9 shows the data has been successfully shuffled. After shuffling, the “Algorithm menu” displaying four options is displayed. The four displayed algorithms are *KMeans*, *DBSCAN*, *Isolation Forest*, and *Local Factor Outlier*.

---

**Fig 9.** Figure displays the “Algorithm menu” displaying four options

Fig. 10 shows option 1 referring to “KMeans” as the selected option. After selecting “KMeans”, the user is asked to enter the number of clusters.
**Fig. 10.** Figure asks to enter the number of clusters after selecting “KMeans”

Fig. 11 shows the entered number of clusters followed by the “Initialization method” as “random” out of the two available options.

**Fig. 11.** Figure shows the selected “Initialization method” as “random” out of the two available options

Fig. 12 shows the successfully clustered data.
Run Time -> --- 0.0 seconds ---
Data Successfully Clustered

KMEANS RESULTS

Clusters -> [0, 1, 2, 3]

Inertia -> 191391.34258222632

<table>
<thead>
<tr>
<th>col_0</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>row_0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>18714</td>
<td>7974</td>
<td>13840</td>
<td>26814</td>
</tr>
<tr>
<td>1</td>
<td>16296</td>
<td>7123</td>
<td>11931</td>
<td>23280</td>
</tr>
</tbody>
</table>

Max True Label

**Fig. 12.** Figure shows the successfully clustered data with the value of inertia

Fig. 13 shows the “KMeans Score metrics Menu” comprising three options.

<table>
<thead>
<tr>
<th>col_0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
</tbody>
</table>

dtype: int64

**Fig. 13.** Figure shows the “KMeans Score metrics Menu”

Fig. 14 shows option 1 has been entered as a preferred choice. Option 1 signifies the $F1$ Score as a chosen metric.
Fig 14. Figure shows option 1 has been entered as a preferred choice

Fig. 15 asks for entering the option from the “KMeans Score Metrics Menu”. The average method is to be selected among the four available methods (weighted, micro, macro, binary).

Fig 15. Figure depicts the technique asking for entering the option from the “KMeans Score Metrics Menu”

Fig 16 shows the obtained results after matching the clusters with the maximum intersection.

Fig 16. Figure shows the obtained results after matching the clusters with a maximum intersection

Case 2:

The case opts for CIC-IDS 2017 dataset. Fig. 17 shows the Manage Missing Values menu with six options: eliminating missing values, imputing 0 for missing values, imputing mean for missing values, imputing median for missing values, imputing mode for missing values, and simple imputer. Option 1 for eliminating the missing values has been chosen and none of the columns was found with missing values.
Fig 17. Manage Missing Values menu

Fig. 18 shows the Encoding Menu. The option selected for encoding is ‘Binary true labels (abnormal)’.

Fig 18. Displays the Encoding Menu

Fig. 19 shows the successful scaling and shuffling of data conducted.
Fig 19. Displays the successful scaling and shuffling of data

Fig. 20 shows the ‘Algorithm Menu’ displaying four algorithms for performing clustering. Option 3 for ‘Isolation Forest’ is chosen with a contamination factor of 0.1 for performing clustering.

Fig 20. Displays the Algorithm menu displaying four algorithms for performing clustering

Fig. 21 shows the time consumed for performing clustering to be 10.582605361 seconds.
Fig 21. Time consumed for performing clustering

Fig. 22 displays the results obtained after the successful execution of clustering.

Fig 22. Displays the results obtained after the successful execution of clustering

Fig. 23 shows the ‘Isolation Forest Score Metrics Menu’ for calculating the F1 score using ‘weighted’ as the ‘Average method’.
Network Intrusion Detection: Comparative Analysis of NSL-KDD and CIC-IDS2017 Datasets

Fig. 23. Displays the ‘Isolation Forest Score Metrics Menu’

Fig. 24. Displays the calculated F1 score of 0.71296240.

5. CONCLUSION
This section elaborates on the results obtained from the two different cases executed. Different encoding techniques have been implemented on the NSL-KDD and CIC-IDS2017 datasets. The research highlights the importance of using a diverse and realistic datasets like NSL-KDD and CIC-IDS 2017 for evaluating intrusion detection models. The dataset's inclusion of various attack scenarios and labeled instances provides a valuable benchmark for assessing the performance of different techniques and comparing their effectiveness.

REFERENCES

Network Intrusion Detection: Comparative Analysis of NSL-KDD and CIC-IDS2017 Datasets


AUTHORS’ BIOGRAPHY

Mr. Ritesh Kumar, is currently serving in the scientific cadre in the Standardisation Testing and Quality Certification, Ministry of Electronics and Information Technology (MEITY), Government of India. His total research and teaching experience is more than 10 years. He has done his B.Tech in Computer Science and Engineering from Dr. B.R. Ambedkar National Institute of Technology, Jalandhar, Punjab, India. He is an M.Tech. research scholar in Computer Engineering from Yadavindra College of Engineering, Punjabi University Guru Kashi Campus, Talwandi Sabo, Punjab, India.

Dr. Gagandeep Jagdev, is currently serving in the capacity of Technical Officer at Punjabi University Guru Kashi Campus, Damdama Sahib (PB). His total teaching and research experience is more than 16 years and has 158 International and National publications in reputed journals and conferences to his credit. He is the guest editor of the Journal of Imaging, MDPI. He is also a member of the editorial board of several reputed International Journals indexed in ESCI, Scopus, ACM, WoS, and Pubmeds and has been an active Technical Program Committee member of several International and National conferences conducted by renowned universities and academic institutions. He has been bestowed with Best Research Paper awards 4 times by NITTTR (Chandigarh), Government College of Engineering and Technology (Jammu), and Guru Nanak College, Budhlada (PB). He has actively participated in more than 80 Webinars and FDPs. His field of expertise is Image Processing, Big Data Analytics, Data Science, Cloud Computing, Cloud Security, Cryptography, and WANETs.