International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 04 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

Performance Evaluation of Dimensionality Reduction Techniques on Big Data using Machine Learning Algorithms

Venkateswara Gupta Pola[1], Pranahith Babu Yarra[2], Ramya. A[3], Sai Suraj Karra[4]

[1]Venkateswara Gupta Pola
Dept. of Computer Science Engineering, SASTRA University

[2]Pranahith Babu Yarra
Dept. of Mechanical Engineering (Mechatronics), Mahatma Gandhi Institute of Technology

[3]Ramya. A
Dept. of Computer Science Engineering, Marri Laxman Reddy Institute of Technology and Management

[4]Sai Suraj Karra
Dept. of Mechanical Engineering (Mechatronics), Mahatma Gandhi Institute of Technology

---

Abstract - *Large volumes of data are generated due to digitization which in-turn created huge datasets. To find specific patterns hidden in the attributes of these datasets, several ML algorithms are used. But in these large datasets, certain attributes may not be helpful or may not affect the prediction result. Ignoring irrelevant attributes reduces burden on ML algorithms and decreases cost and computation power which can be done using dimensionality reduction techniques.*

In this project two prominent Dimensionality Reduction algorithms are used namely Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) on Intrusion detection dataset with ML Techniques like DT, NB, SVM and RF classifiers. The project was completed with:

1. *4 Machine learning algorithms namely DT, SVM, NB, RF with two Dimensionality reduction algorithms PCA and LDA respectively.*
2. *PySpark which is a Big Data tool used in this project for implementing the machine learning and dimensionality reduction algorithms.*
3. *Encoders like Label Encoder and One Hot Encoder for converting categorical data to binary data making it suitable for ML algorithms.*

***Key Words*: Machine learning, PCA, LDA, Algorithms.**

1. INTRODUCTION

In the past 15 years, Machine Learning has been one among the fastest growing technologies. It has various applications in different fields, such as computer vision, computational biology, data science, healthcare, banking, criminal identification, market prediction, etc. ML enables a computing machine to learn from a huge sample of information and predict the trends within the raw data. In various research areas, ML algorithms are used to predict hidden patterns and identify the test data to generate accurate results. The scope of ML classification models in the area of cybersecurity is constantly growing. They have proven to be very effective in detecting numerous threats and malicious content.

1.1 WHAT IS INTRUSION DETECTION?

We are being reliant day by day on networks and computer technologies. This raises the need for stable networks. For data privacy, confidentiality and availability, we have to strengthen computer network protection. But these do not stop the detection of intrusion. It is important to protect insecure computing systems and networks only to avoid the possibility of unwanted access and data theft. An Intrusion Detection System checks all the network packets and tries to categorize the traffic as disruptive or non-intrusive. The identification of intrusion is the mechanism that starts where the firewall stops.

1.2 WHY DIMENSIONALITY REDUCTION?

Raw Datasets needs to be pre-processed for any type of Intrusion Detection. In this context, reduction of number of dimensions in the high dimensional data is very important. Numerous dimensionality reduction techniques have been used in the last few decades to filter the data samples of the dataset considered. Dimensionality reduction involves the mapping of high-dimensional to lower-dimensionality inputs such that identical points are mapped to neighboring points on the n- manifold in the input space. Dimensionality Reduction de- creases a huge data processing overhead on the Machine Learning algorithms and decreases the chances of model being overfitted. This also increases the performance of the algorithm at a very significant rate.

2. OBJECTIVE

Analyze the performance of Machine Learning models with and without Dimensionality Reduction. Reduction is done using two major techniques namely PCA and LDA. Raw Intrusion Detection Data is transformed and decomposed using Label and One Hot Encoders, then this undergoes dimensionality reduction process. The reduced features are then fed to the ML Techniques like DT, NB, SVM and RF.

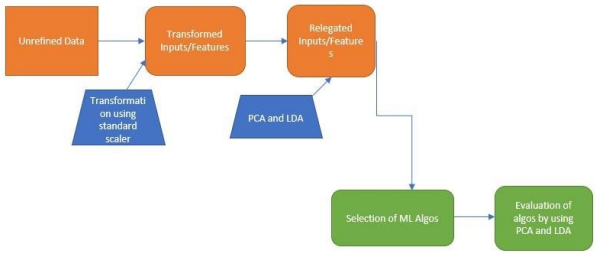

Fig -1: Evaluation of Dimensionality Reduction

In the above figure illustrates the process how the performance evaluation of Dimensionality reduction is done. First, we take the necessary data attributes(features) and labels (customer transactional data) from the dataset and encode categorial data to numerical encoded data. Then perform PCA and LDA to decrease dimensionality of the data.

3. METHODOLOGY

In this project we have used two popular DR algorithms namely Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) with ML techniques like DT, NB, SVM and RF. Dataset undergoes Data pre-processing, Feature engineering and then application of various machine learning algorithms.

1.1 DATA PRE-PROCESSING

In this process two popular encoding techniques are used namely One-Hot Encoding and Label Encoding. Both these methods are used to deal with categorical data but in a different way. One-Hot encoders simply create a new feature in the Dataset while Label Encoders label them with numerical values.

1.2 FEATURE ENGINEERING

Feature engineering is a very important step in pre-processing of data that aims to remove transformed features from the raw dataset, simplifying the ML model and enhancing the consistency of a machine learning algorithm's output. Machine Learning professionals invest much of their time in data cleaning and engineering functionality. The two techniques used are PCA and LDA.

1.3 PRINCIPAL COMPONENT ANALYSIS

This is a mathematical model that uses orthogonal conversion. A group of correlated variables is transformed by PCA to a group of uncorrelated variables. For exploratory data processing, PCA is used. For analysis of the relationships between a group of variables, PCA may also be used. It can thus be used to minimize dimensionality.

Let us consider a n-dimensional input.

$$x^{(1)}, x^{(2)}, x^{(3)}, \ldots\ldots, x^{(m)}$$

The dimensionality is reduced to k -dimensions where $k \ll n$ This is done using following these steps of PCA.

1. Standardization of Data: The data fed must have zero mean and unit variance.

$$x_j^i = \frac{x_j^i - \bar{x_j}}{\sigma_j} \quad \forall_j$$

2. Calculation of the co-variance matrix for the given data.

$$\sum = \frac{1}{m} \sum_i^m (x_i)\,(x_i)^T, \sum \in R^{n*n}$$

3. Calculation of Eigen Vectors and Eigen Values

$$u^T \sum = \gamma\mu$$

$$U = \begin{matrix} | & | & | \\ u_1 & u_2 \ldots . & u_n \\ | & | & | \end{matrix}, u_i \in R^n$$

4. Projection of given raw data as a k-dimensional sub-space Now we can choose top k eigen vectors from this matrix formed.

$$x_i^{new} = \begin{matrix} u_1^T x^i \\ u_2^T x^i \\ ::::: \\ ::::: \\ u_k^T x^i \end{matrix} \in R^k$$

1.4 LINEAR DISCRIMINANT ANALYSIS

Another prominent technique for dimensionality reduction is LDA. LDA focuses on reduction of high dimensional data into lower dimensions with a good class level separability which will indeed reduce computation costs. Process of LDA is similar to PCA where PCA uses technique of maximizing variance while LDA maximizes separation of classes. LDA ensures reduction of higher dimensional data into lower subspace say i (where $i \leq x - 1$)) without any disturbance in information of class.

LDA is generally performed in 5 steps:

1. Computing a m-dimensional mean vector for every class in the given dataset.
2. Carrying out scattered matrix computation.
3. Computing eigen vectors $(E_1, E_2, E_3, \ldots\ldots, E_m)$ and their eigen values $(\varphi_1, \varphi_2, \varphi_3, \ldots\ldots, \varphi_m)$ for the computed scatter matrices.
4. Opting k eigen vectors having maximum eigen values by arranging them in descending order, forming a m * i matrix ZZ.
5. Transforming the input data into new subspace using m * i matrix.

OO = XX * ZZ

1.5 APPLYING MACHINE LEARNING TECHNIQUE

Machine learning algorithms are now applied on the output dataset produced after dimensionality reduction is applied. The algorithms proposed are applied with PCA, with LDA and without any DR. Then the results are analyzed.

4. DATASET

The dataset used for this project is a standard IDS dataset NSL-KDD that is an improved version of famous KDD99 dataset. This dataset has a huge volume of attacks and attack parameters.

```
0,tcp,private,REJ,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,229,10,0.00,0.00,1.00,1.00,0.04,0.06,0.00,255,
0,tcp,private,REJ,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,136,1,0.00,0.00,1.00,1.00,0.01,0.06,0.00,255,1
2,tcp,ftp_data,SF,12983,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,134
0,icmp,eco_i,SF,20,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,65,0.00,0.00,0.00,0.00,1.00,0.00,1.00,3,57,1.
1,tcp,telnet,RSTO,0,15,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,8,0.00,0.12,1.00,0.50,1.00,0.00,0.75,29,86,
0,tcp,http,SF,267,14515,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,4,4,0.00,0.00,0.00,0.00,1.00,0.00,0.00,155,2
0,tcp,smtp,SF,1022,387,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,3,0.00,0.00,0.00,0.00,1.00,0.00,1.00,255,28
0,tcp,telnet,SF,129,174,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,255,2
0,tcp,http,SF,327,467,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,33,47,0.00,0.00,0.00,0.00,1.00,0.00,0.04,151,2
0,tcp,ftp,SF,26,157,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,52,26,0.5
0,tcp,telnet,SF,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,255,128,0
0,tcp,smtp,SF,616,330,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,2,0.00,0.00,0.00,0.00,1.00,0.00,1.00,255,129
0,tcp,private,REJ,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,111,2,0.00,0.00,1.00,1.00,0.02,0.07,0.00,255,2
0,tcp,telnet,S0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,120,120,1.00,1.00,0.00,0.00,1.00,0.00,0.00,235,1
37,tcp,telnet,SF,773,364200,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,3
0,tcp,http,SF,350,3610,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,8,8,0.00,0.00,0.00,0.00,1.00,0.00,0.00,71,255
0,tcp,http,SF,213,659,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,24,24,0.00,0.00,0.00,0.00,1.00,0.00,0.00,255,2
0,tcp,http,SF,246,2090,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,16,16,0.00,0.00,0.00,0.00,1.00,0.00,0.00,35,2
0,udp,private,SF,45,44,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,505,505,0.00,0.00,0.00,0.00,1.00,0.00,0.00,25
0,tcp,private,REJ,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,204,18,0.00,0.00,1.00,1.00,0.09,0.07,0.00,255,
0,tcp,ldap,REJ,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,118,19,0.00,0.00,1.00,1.00,0.16,0.05,0.00,255,19,
0,tcp,pop_3,S0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1.00,1.00,0.00,0.00,1.00,0.00,0.00,255,87,0.3
0,tcp,http,SF,196,1823,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,17,17,0.00,0.00,0.00,0.00,1.00,0.00,0.00,255,
0,tcp,http,SF,277,1816,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,17,18,0.00,0.00,0.00,0.00,1.00,0.00,0.11,36,2
0,tcp,courier,REJ,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,116,8,0.00,0.00,1.00,1.00,0.07,0.07,0.00,255,8
0,tcp,discard,RSTO,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,273,13,0.00,0.00,1.00,1.00,0.05,0.06,0.00,255
0,tcp,http,SF,294,6442,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,22,46,0.00,0.00,0.00,0.00,1.00,0.00,0.11,180,
0,tcp,http,SF,300,440,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,7,7,0.00,0.00,0.00,0.00,1.00,0.00,0.00,255,255
0,icmp,ecr_i,SF,520,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,511,511,0.00,0.00,0.00,0.00,1.00,0.00,0.00,46,
0,udp,private,SF,54,51,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,511,511,0.00,0.00,0.00,0.00,1.00,0.00,0.00,25
805,tcp,http,RSTR,76944,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,12,12,0.00,0.00,1.00,1.00,1.00,0.00,0.00,2
0,tcp,smtp,SF,720,281,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,2,0.00,0.00,0.00,0.00,1.00,0.00,1.00,158,82,
0,tcp,http,SF,301,19794,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,15,15,0.00,0.00,0.00,0.00,1.00,0.00,0.00,20,
0,udp,private,SF,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,40,3,0.00,0.00,0.00,0.00,0.08,0.38,0.00,255,3,0
```

Fig -2: Sample Dataset

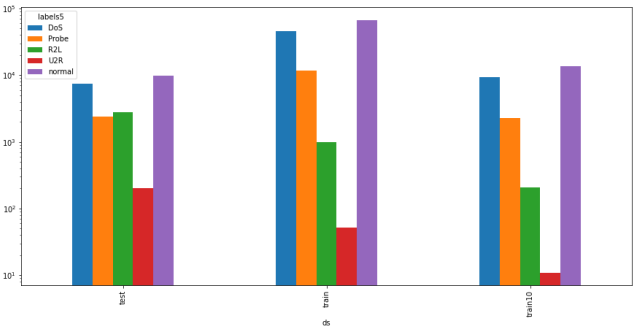

Fig -3: Dataset Attack Types

5. RESULTS

The dataset used for this project is a standard IDS dataset NSL-KDD that is an improved version of famous KDD99 dataset. This dataset has a huge volume of attacks and attack parameters.

S.No	Algorithm Name	Direct	PCA	LDA
1	Random Forest	76.7033	84.3748	83.6876
2	Decision Tree	78.1583	75.5867	83.3848
3	Naive Bayes	75.8029	84.6019	87.1935
4	SVM	75.4258	84.1302	85.7725

Table-1: Final Experimental Results

6. CONCLUSIONS AND FUTURE RESEARCH

According to Table 1 performance of the algorithms is best when Dimensionality reduction techniques are applied. We can observe that there is a drastic improvement in model accuracy. On given dataset (IDS) Naïve Bayes with LDA performs very well with a highest accuracy of 87.1935 %. As a comparative study PCA and LDA both perform equally better while Decision Tree with PCA is not as expected.

In the future, we can extend to assess the efficacy of these dimensionality reduction strategies on high-dimensionality data such as images, BLOBs etc. These approaches can also be used on more complex algorithms such as Deep Neural Networks, Recurrent Neural Networks, Convolution Neural Networks, etc.

REFERENCES

[1] G. Thippa Reddy, M. Praveen Kumar Reddy, Kuruva Lakshmana, Rajesh Kaluri, Dharmendra Singh Rajput, Gautam Srivastava. (2020). "Analysis of Dimensionality.

[2] Priyanka Dahiya., Devesh Kumar Srivastava. (2018). "Network Intrusion Detection in Big Dataset Using Spark" International Conference on Computational Intelligence and Data Science.

BIOGRAPHIES

Venkateswara Gupta Pola pursuing Bachelor of Technology in Computer Science and Engineering at SASTRA University, Thanjavur and working on VQA and Deep learning Research projects at University Autonoma De Barcelona, Spain.

Pranahith Babu Yarra pursuing Bachelor of Technology in Mechatronics Engineering at Mahatma Gandhi Institute of Technology, Hyderabad, currently doing Project on Automatic CNC Laser cutting in MGIT.

Ramya. A pursuing Bachelor of Technology in Computer Science and Engineering at Marri Laxman Reddy Institute of Technology, Hyderabad, currently working on YOLO based blood cells detection in MLRIT.

Sai Suraj Karra pursuing Bachelor of Technology in Mechatronics Engineering at Mahatma Gandhi Institute of Technology, Hyderabad, currently doing Project on Automatic CNC Laser cutting in MGIT.