1 Introduction

1.1 Introduction

Many data mining techniques exist like classification, clustering, association rule. The of data is time consuming but it identifies complex type of IDS data. Association rule technique is based on attribute/ value pair. It identifies multi-feature correlation between attributes of IDS and based on this correlation classification of attack is done. Support vector machine (SVMs) was first proposed by Vapnik [19] which is used for binary classification. SVM prove that SVM is fit for the highly variable and high dimensional data. Data acquired in IDS domain is more complex. So we introduce SVM to the research for intrusion detection system (IDS). SVM is a classifier which can be store the result in two ways: partial or true. Generally, SVM is a simplest linear form

Support Vector Machine (SVM) Algorithm is specially used for solving the classification problem. The fundamental concept of SVM is to find decision surface that separates the best data vectors into two classes for a given problem which is defined over the vector space SVM is a simplest linear form generally it is a hyperplane that separates the positive examples from the negative example. Figure. 1 shows the hyperplane that separates the training data by a maximal margin between two classes [20]. All vectors lies on one side of the hyperplane labeled as +1 and other vectors lies on another side of the hyperplane labeled as -1. The training objects are lying closest to the hyperplane are called the support vectors

1.2 Problem statement:

Classification of KDDCup dataset using SVM Algorithm with Elastic net regularization.

1.3 Objective

* The objective of this project is used to classify the KDDCup dataset using svm with elastic net regularization to optimise the performance of the algorithm
* Enhance the time complexity of the model for KDDCup dataset.

1.4 Scope

• Increase the scalability performance of the algorithm for the large dataset

• Enhances complexity of large dataset

2 Literature survey

2.1 Existing system

  Classification of KDDCup dataset using Support Vector Machines with with Elastic net algorithm ... The past years have witnessed many dedicated open-source projects that built and maintain ... (SVM), parallelized for GPU, multi-core CPUs and distributed systems. Elastic Net naturally takes advantage of the vast existing work on ...

2.2 Proposed system:

The experiments on KDDCUP99 dataset using svm and elastic net regulizers Classification is done using SVM classifier. It gives the result either in “One” or “zero” results. During classification of attacks, any type of attack is classified properly to their belonging class then “One” result is displayed and accuracy of classification is increases as per the “Zero” result. If any type of attack is not classified to their belonging class means it appears in wrong class, then the result shown “Zero”.

3 Specific Requirements

* 1. Functional Requirements

Introduction: - In this project is used to classify the KDDCup dataset using svm with elastic net regularization to optimise the performance of the algorithm

Inputs:-In our project we are using Bench Mark Data Set KDDCup data set as input

Processing: - One-Hot-Encoding is used to transform all categorical features into

binary features. The One-Hot-encoding takes a matrix of integers, denoting the values on by categorical features. The output will be a sparse matrix where each column corresponds to one possible value of one feature. Therefore the features first need to be transformed with Label Encoder, to transform every category to a number. In the second step, we will be applying various machine and deep learning algorithms. These algorithms analyse the large datasets and mechanism which show the intrusion in the given KDDCup dataset with accuracies.

Output:

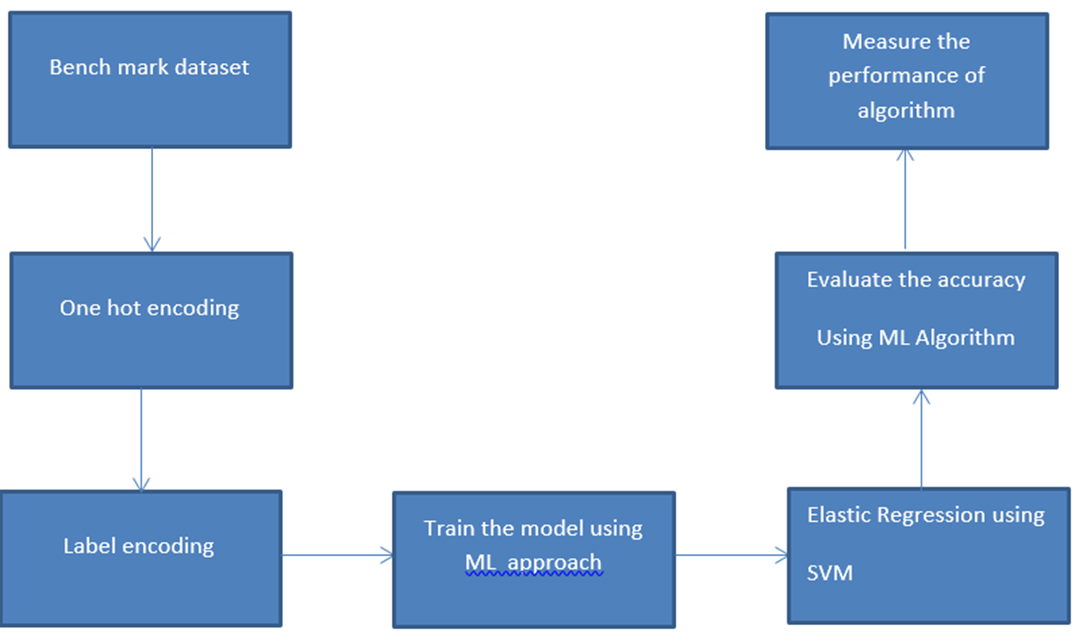
SVM-

3.2 Non-functional requirements

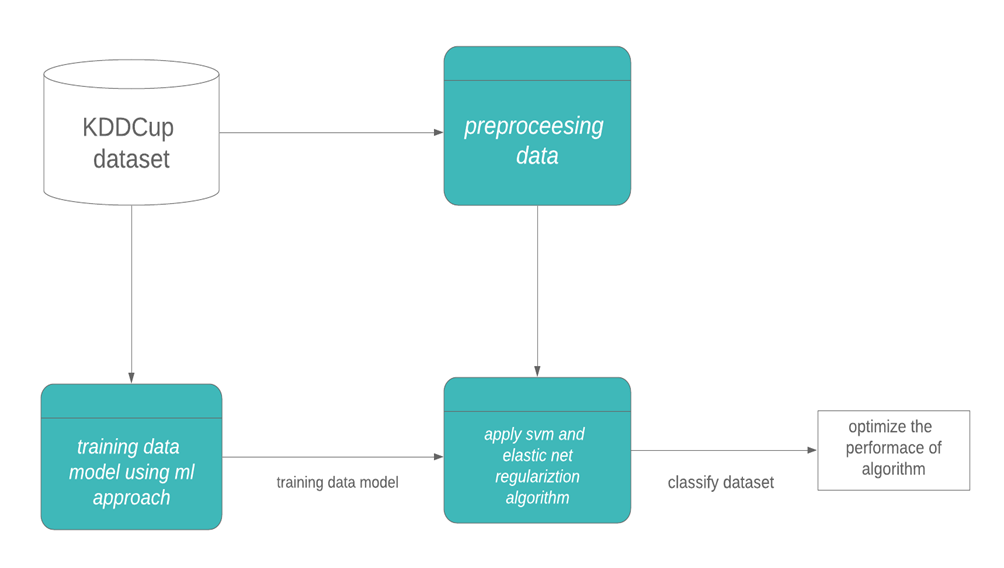
|  |  |
| --- | --- |
| Type | Description |
| Performance | 1.The system should be able to classify anomalies and normal packets with the accuracy of more than 95%.  2. It get accuracy |
| Usability | 1.The system should be available all the time. |

4 System Design

4.1 Block diagram



4.2 Data flow diagram



1. Data-set Contents:-

|  |  |  |
| --- | --- | --- |
| Feature  No. | Feature Name | Description |
| 1. | Count | No. of connections to the same host as the current connection in the last two seconds |
| 2. | destination bytes | Bytes sent from destination to source |
| 3. | diff srv rate | percentage of connections to different  services |
| 4. | dst host count | count of connections having the same  destination hosts |
| 5. | dst host diff srv rate | percentage of different services on the current host |
| 6. | dst host rerror rate | percentage of connections to the current host that has an RST error |
| 7. | dst host same src port rate | percentage of connections to the current host having the same src port |
| 8. | dst host same srv rate | percentage of connections having the  same destination host and using the same service |
| 9. | dst host serror rate | percentage of connections to the current host that have an S0 error |
| 10. | dst host srv count | count of connections having the same  destination host and using the same service |
| 11. | dst host srv diff host rate | percentage of connections to the same  service coming from different hosts |
| 12. | Dist host srv rerror rate | percentage of connections to the current host and specified service that have an RST error |
| 13. | dst host srv serror rate | percentage of connections to the current host and specified service that  have an S0 error |
| 14. | Duration | Duration of the active connection. |
| 15. | Flag | Status flag of the connection |
| 16 | Hot | No. of "hot" indicators |
| 17. | is guest login | One if the login is a "guest.''  login; Otherwise 0 |
| 18. | is host login | One if the login belongs to  the "host''; otherwise 0 |
| 19. | Land | One if the connection is  from/to the  same host/port;  Otherwise 0 |
| 20. | logged in | One if successfully logged  in; otherwise 0 |
| 21. | num access files | No. of operations on  access control files |
| 22. | num compromised | No. of compromised  conditions |
| 23 | num failed logins | No. of failed logins |
| 24. | num file creations | No. of file creation  operations |
| 25. | num outbound cmds | No. of outbound  commands in an ftp  session |
| 26. | num root | No. of "root'' accesses |
| 27. | num shells | No. of shell prompts |
| 28. | protocol type | Connection protocol  (e.g. tcp, udp). |
| 29. | rerror rate | percentage of  connections that have  “REJ'' Errors |
| 30. | root shell | One if the root shell is  obtained; otherwise 0 |
| 31. | same srv rate | percentage of  connections to the same  service |
| 32. | serror rate | percentage of  connections that have  “SYN'' Errors |
| 33. | Service | Destination service (e.g.  telnet, ftp) |
| 34. | src bytes | Bytes sent from source to destination |
| 35. | srv count | No. of connections to the same service as the current connection in the last two seconds |
| 36. | srv diff host rate | percentage of connections to different  hosts |
| 37. | srv rerror rate | percentage of connections that have  “REJ'' errors |
| 38. | srv serror rate | percentage of connections that have  “SYN'' Errors |
| 39. | su attempted | One if "su root'' command attempted; otherwise 0 |
| 40. | Urgent | No. of urgent packets |
| 41. | Wrong fragment | No. of wrong fragments |

6. Implementation

6.1 Introduction

Anaconda

Anaconda is a freemium open source distribution of the Python and R programming languages for large-scale data processing, predictive analytics, and scientific computing, that aims to simplify package management and deployment. Its package management system is conda. Anaconda distribution comes with more than 1,000 data packages as well as the Conda package and virtual environment manager, called Anaconda Navigatorso it eliminates the need to learn to install each library independently.

Jupyter

Jupyter Notebook (formerly IPython Notebooks) is a web-based interactive computational environment for creating Jupyter notebooks documents. The "notebook" term can colloquially make reference to many different entities, mainly the Jupyter web application, Jupyter Python web server, or Jupyter document format depending on context.

Python

Python is an interpreted, high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales. In July 2018, Van Rossum stepped down as the leader in the language community. Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms. It includes object oriented, imperative, functional and procedural, and has a large and comprehensive standard library. Python interpreters are available for many operating systems. CPython, the reference implementation of Python, is open source software and has a community-based development model, as do nearly all

of Python's other implementations. Python and CPython are managed by the non-profit Python Software Foundation.

6.2 Algorithms used for Implementation

Support vector machine:-

SVM is a supervised machine learning algorithm which can be used for classification or regression problems. It uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs.

7. Result and analysis

* We used KDDCup dataset which contains 42 features in which we have extracted 7 features using confusion matrix. We applied the SVM algorithm for processed dataset.
* Training: 70%
* Testing: 30%

**9** Conclusion and future work

* The DDoS attacks in a network is a challenging problem as it involves recording the log old flow traffic and detection of DsDos attacks from multiple authentic flow traffic.
* SVM is a statistical learning theory. This Theory characterised the performance of learning machine using bonds on their ability to predict feature data.