# A Review on Big Data Classification: using Machine Learning Technique to Classify Intrusion

[1]Nilamadhab Mishra, [2]Sarojananda Mishra
[1]Research Scholar , [2]Professor and Head
[1]Computer Science and Engineering
[1]Biju patnaik University of Technology, Rourkela, Odisha, India
[2]Computer Science and Engineering
[2]IGIT, Sarang, Odisha, India

***Abstract:***

Big Data Analytics is anactivity of examine and accepting the singularity and features of very big size datasets by retrieving useful numerical and statistical patterns. It is a composite process of probing big and diversity data sets or big data to uncover information. These types of dataset increase the complexity of the data and thus make the current techniques and technologies stop working as expected within a given process time. Many applications suffer from the Big Data difficulty, including network traffic risk analysis, geospatial classification and business forecasting. Network intrusion finding and prediction are time sensitive applications and they need highly efficient Big Data techniques and technologies to handle the problem. The modern technologies can help to solve Big Data analytics on various applications. The troubles and challenge associated with the modern networking technologies and machine learning techniques can be used to solve Big Data classification problem for network intrusion forecast.

***Keywords:*****Big Data, Data Mining, Intrusion detection, Machine Learning.**

## I. Introduction

In the present situation, it is difficult to work without internet. Every person has addiction on internet. It has become an significant model in various applications such as learning, business and others. So security of the data that is communicated through internet is necessary. Securenetwork is maintained by Intrusion Detection System (IDS). IDS observe the data interchange carefully and identify it as normal or spam. Nowaday's most of the applications depends on the advance network technologies namely wireless networks, wireless sensor networks and blue tooth. In case of wireless sensor networks security mechanisms such as key-management protocols, authentication techniques and security protocols cannot be used because of resource constraints. Intrusion Detection System is the ideal security method for wireless sensor networks. [1]

Usual Intrusion prevention technique such as firewall, access control and encryption has failed to detect the intrusion in the networks. As a result Intrusion detection system becomes an essential component. The idea of the Intrusion detection system (IDS) is to prevent the computer system from attack. The IDS is the most essential part of the security infrastructure for the networks connected to the internet because various ways to compromise the stability and security of network.

IDS can be classified into two types: Anomaly and Misuse detection. Anomaly detection system creates a database of normal behavior and any deviations from the normal behavior are occurred an alert is triggered regarding the occurrence of intrusions. Misuse Detection system stores the Predefined attack patterns in the database if a similar data and if similar situations occur, it is classified as attack. Based on the source of data the intrusion detection system are classified to Host based IDS and Network based IDS. In network based IDS the individual packet flowing through the network are analyzed. The host based IDS analyzes the activities on the single computer or host. The main disadvantage of the misuse detection method is that it cannot detect novel attacks and variation of known attacks. To avoid these drawbacks we go for anomaly based detection methods. With this approach, known and novel attacks can be detected. The problem is that it will generate more false alarms. The intrusion detection method based on unsupervised learning has a high detection rate but also a high False positive rate [2].

Intrusion detection functions include:-

Monitoring and examining of both user and system activities.

Analyzing system configurations and weaknesses.

Assessing system and file integrity.

Skill to identify patterns typical of attacks.

Analysis of irregular activity patterns.

Tracking user policy violations.

Based on the literature study the number of intrusion detection systems are developed using different machine learning techniques. Some study apply single learning techniques as individual; some systems are combining two or more different learning techniques, such as fusion techniques, and combining multiple weak learners to improve the concert of a classifier known as all together classifier. Particularly these techniques are developed as classifiers and clusters. The classifiers techniques are supervised learning that classify or recognize whether the internet activity is normal or attack, and cluster techniques are unsupervised learning that is trying to find a covered structure of unlabeled cluster data.

So the aim of this paper is to review related studies published in the past decade by investigative the techniques that have been used. Different experiments have been conducted, based on the machine learning algorithm viewpoint what should be considered for future work.

## II. Big Data

Big data is a field which based on the ways to examine, retrieved information from, or deal with data sets which are too large or compound to be handled with by traditional data giving out application software. Data with many cases offer greater statistical power, while data with higher difficulty may lead to a higher false discovery rate. Big data challenge includes capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy and data source. Big data was associated with three key concepts: volume, variety, and velocity. But now another two points added i.e. veracity and value. When we handle big data, we must have to observe and track what happens. Thus, big data often includes data and sizes that exceed the capacity of traditional usual software for processing within an acceptable value and time [3].

**Types of Big Data:**

Big Data could be obtainable in three forms:

- Structured
- Unstructured
- Semi-structured

**Structured:**

Several kinds of data that can be stored, accessed and processed in the form of a fixed format is termed as structured data. Over the period of time, talent in computer science has achieved the greater success in developing techniques for working with such kind of data and also generating value out of it. However, nowadays, we are forecast issues when a size of such data grows to a enormous extent, classic sizes are being in the range of multiple zetabytes.

**Unstructured:**

Several kind of data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, unstructured data poses different challenge in terms of its processing for generating value out of it. A typical example of unstructured data is a varied data source containing an addition of simple text files, images, videos etc. Now a day's organizations have wealth of data available with them but unfortunately, they don't know how to derive value from it, since this data is in its unrefined form or unstructured format.

**Semi-structured:**

Semi-structured data contain both the form of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational database management system. Example of this type of data is a data, represented in an extensible markup language file (XML).

**Characteristics of Big Data:**

(i)Volume – Big Data is related to a size which is huge. Size of data plays a very important role in formative value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is depend upon the volume of data. Hence, Volume is one characteristic which requires to be considered while trade with Big Data.

(ii) Variety – The next point of Big Data is its variety.

Variety refers to different source and the actions of data, both structured and unstructured. During the old days, excel sheets and databases were the only sources of data considered by most of the applications. Nowadays, data is in the form of photos, emails, videos, monitoring devices, audio, PDFs, etc. are also being considered in the analysis of applications. This mixture of unstructured data contains particular issue for data storage, data taking out and analyze of data.

(iii) Velocity – Velocity refers to the speed of generating the data. How fast the data is being generated and processed to meet the requirements, determines actual potential in the data.

Big Data Velocity deals with the speed from end to end which is the data flows from sources like business processes, application logs, networks, and social media sites, sensors, Mobile devices, etc. The flow of data is enormous and uninterrupted.

(iv) Variability/Veracity – Veracity is the changeableness which can be detected by the data at times, thus inhibits the process is being able to handle and direct the data effectively.

(v)Value – Extract useful data as per our requirements.

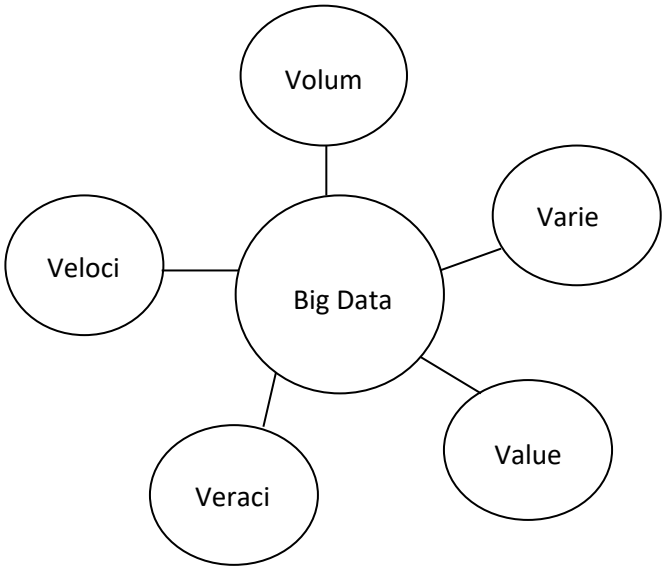


**Fig-1 Big Data Characteristics**

**Profit of Big Data Processing:**

Big Data gives the following benefits:

- Businesses can make use of outside intelligence while taking decisions.

Access to social data from search engines and sites like twitter, facebook are enable organization to refresh their business strategy.

- Developed the customer service.

Usual customer feedback systems are replace by new systems designed with Big Data technologies. In the new systems, Big Data and natural language processing technologies both are being used to read and evaluate consumer experience.

- Before time identification of risk to the product or services, if any.
- Enhanced operational effectiveness.

Big Data technologies can be used for creating a particular area or corridor zone for new data before identifying what type of data should be moved to the data storehouse. Such mixing of Big Data and data warehouse technologies makes an organization to take down uncommonly accessed data.

## III. Data Mining

Data mining is used for unseen, suitable, and potentially useful patterns in big data sets. Data mining in general discover unsuspected or formerly unknown relationships among the data. It is a multi-disciplinary skill that uses machine learning, statistics and AI and database technology. That means derived via Data Mining can be used for marketing, scam detection, and methodical discovery etc.
Data mining is otherwise called as Knowledge detection, Knowledge mining, data/pattern analysis, information harvesting etc.

**Type of Data:**

Data mining can be performed by following types of data

- Relational database.
- Data warehouse.
- Advanced DB and information repository.
- Object-oriented and object-relational database.
- Transactional and Spatial database.
- Heterogeneous and heritage database.
- Multimedia and streaming database.
- Text database.
- Text mining and Web mining.

**Data Mining Technique:**

**1. Classification:**

This study is used to extract important and applicable information about data, and metadata. This method helps to classify data in multiplicity classes.

**2. Clustering:**

This analysis is a data mining method to identify data that are like each other. This procedure helps to detect the difference and similarity between the data.

**3. Regression:**

This study is the data mining method of identify and analyze the rapport between variables. It is used to identify the similarity of a specific variable, given the occurrence of other variables.

**4. Association Rules:**
This technique helps to find the grouping between two or more items. It discovers a unseen pattern in the data set.
**5. Outer detection:**
This data mining technique refers to the watching of data items in the dataset which is not matching an expected pattern or expected behavior. This technique also can be used in different areas, such as intrusion detection, scam or fault detection, etc. This is also called Outlier Analysis or Outlier mining.
**6. Sequential Patterns:**
This help to discover or identify related patterns or trends in deal data for certain period.
**7. Prediction:**
This can be used as a combination of the other data mining techniques like trends, sequential patterns, clustering, classification etc. This analyzes past events or instance in a perfect sequence for predicting a future event.

## IV. MACHINE LEARNING TECHNIQUES

Machine learning is a division of Artificial Intelligence that provides the systemability to learn automatically and improve from knowledge without being unequivocally programmed. These techniques are used to know the pattern. Pattern Recognition is a task to take raw data and action on data category identification. Supervised and unsupervised algorithms are used to solve different pattern recognition problems. Supervised learning is an assignment infers a function from training labeled data, in which each training data contains a pair of an input vector and class label. Unsupervised culture is used to draw inferences from input dataset without consisting class label.

***Single Classifiers***

IDS model is building up using one single machine learning algorithm. From literature study, the following different machine learning algorithms are used to solve the problems.

- K-nearest neighbor
- Support vector machine
- Decision trees
- Artificial neural network
- Self-organizing maps
- Genetic Algorithms
- Naïve Bayes
- Fuzzy logic
- Hybrid Classifiers
- Ensemble Classifiers

## V. CONCLUSION

In present scenarios, application supervision and mining Big Data is a big tough task. The latest representation or learning technique and support vector machine is to forecast network intrusions through Big Data classification strategy. It is suggested that machine learning structure for solving the problems related with the link parameters. It is also discussed the problems and challenges that the Big Data classification system for network intrusion prediction have to experience during the Big Data analytics. Study on Big Data techniques and technologies develop and at the same time new problems and challenges are rising, hence, the hope is to expand better techniques and technologies towards verdict solutions for Big Data classification difficulties.

## REFERENCES:


**[1]** https://www.scribd.com/document/403498088/Performance-Evaluation-of-Supervised-Machine-Learning-Algorithms-for-Intrusion- Detection.
**[2]** https://www.researchgate.net/scientific-contributions/75085901_Zou_Xinguo
**[3]** https://en.wikipedia.org/wiki/Big_data
**[4]** S. Manocha , M.A. Girolami, "An empirical analysis of the probabilistic K-nearest neighbour classifier", Science Direct, Pattern Recognition Letters, 28 (2007), 1818–1824.
**[5]** C.M.Bishop. (1995). Neural networks for pattern recognition. England: Oxford University.
**[6]** Mitchell, T. (1997). Machine learning. New york: MacHraw Hill.
**[8]** J. R. Quinlan, "Introduction of Decision Trees", Machine Learning, Centre for Advanced Computing Sciences, New South Wales Institute of Technology, Sydney ,2007, vol. 1
**[9]** S. Haykin, Neural networks: A comprehensive foundation (2nd ed.), Prentice Hall, New Jersey, U.S.A, 1999.
**[10]** Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43, 59–69.
**[11]** Koza, J. R. (1992). Genetic programming: On the programming of computers by means of natural selection. Massachusetts: MIT.
**[12]** Pearl, J. (1988). Probabilistic reasoning in intelligent systems. Morgan Kaufmann.
**[14]** J. -S. Jang, C. -T. Sun, and E. Mizutani, Neuro-fuzzy and soft computing: A computational approach to learning and machine intelligence. Prentice Hall, New Jersey, U