Mathematical Statistician and Engineering Applications
ISSN: 2326-9865
DOI: https://doi.org/10.17762/msea.v70i1.2300

Bayesian Classification Model for Network Intrusion Detection Using Clustering Analysis

Lakshita Sajal

Asst. Professor, Department of Computer Science, Graphic Era Hill University, Dehradun, Uttarakhand India 248002

Article Info
Page Number: 198-206
Publication Issue:
Vol. 70 No. 1 (2021)

Article History
Article Received: 25 January 2021
Revised: 24 February 2021
Accepted: 15 March 2021

Abstract
In order to construct an IDS that is both computationally effective and efficient, the goal of this work is to pinpoint significant decreased input characteristics. For this, we use information gain, gain ratio, and correlation-based feature selection to examine the effectiveness of three common feature selection techniques. NSL KDD dataset to identify assaults on the four attack types: Probe (information gathering), DoS (denial of service), U2R (user to root), and R2L (remote to local). The signatures of known attacks are often kept in a regularly updated database. It must be educated for new attacks before it can detect them. The goal of anomaly detection is to spot behavior that deviates from the usual. This method revolves around the recognition of unusual traffic patterns. Two methods are frequently used for feature reduction. A Wrapper assesses the value of features using the intended learning method itself, whereas a filter does so using heuristics based on the overall properties of the data.

1. Introduction

Due to the internet's explosive expansion, there are now exponentially more cyber threats including attacks. Network intrusion detection has become a critical aspect of maintaining the security of computer networks. Intrusion detection systems (IDS) are designed to identify suspicious and malicious activities in network traffic and prevent potential attacks. In order to construct an IDS that is both computationally effective and efficient, the goal of this work is to pinpoint significant decreased input characteristics. One approach to intrusion detection is the Bayesian classification model, which is a probabilistic approach to machine learning. The Bayes theorem, which asserts that the likelihood of an event happening, is the foundation of this model based on prior knowledge can be updated with new evidence. In the context of network intrusion detection, the Bayesian classification model can be used to determine the probability that a network traffic pattern represents an attack.

With the objective to establish a Bayesian model for classification for network intrusion detection, it is essential to identify the features that are most relevant for distinguishing between normal and malicious traffic. This process is known as feature selection, and there are various techniques that can be used to identify the most relevant features. In this work, information gain, gain ratio, and correlation-based feature selection are examined to assess the effectiveness of these techniques. The NSL KDD dataset is used to evaluate the performance of the Bayesian classification model. This dataset is a modified version of the KDD Cup 1999 dataset, which was created to evaluate intrusion detection systems. The NSL KDD dataset includes four types

of attacks: Probe (information gathering), DoS (denial of service), and U2R (user to root), and R2L (remote to local). The signatures of known attacks are often kept in a regularly updated database. However, before the IDS can detect new attacks, it must be trained to recognize them.

The goal of anomaly detection is to identify behavior that deviates from the usual. This approach revolves around the recognition of unusual traffic patterns that may indicate a potential attack. Anomaly detection can be used in combination with the Bayesian classification model enhancing intrusion detection's precision. In addition to feature selection and anomaly detection, clustering analysis is another technique that can be used enhancing intrusion detection's precision. Clustering is a method of grouping similar data points together based on their characteristics. In the context of network intrusion detection, clustering can be used to group network traffic patterns into clusters, which can then be used to identify potential attacks.

Two methods are frequently used for feature reduction in clustering analysis. A Wrapper assesses the value of features using the intended learning method itself, whereas a filter does so using heuristics predicated upon the overall qualities of the data. The choice of feature reduction technique depends on the specific requirements of the intrusion detection system and the characteristics of the data. In conclusion, the development of effective intrusion detection systems is critical to maintaining the security of computer networks. The Bayesian classification model, in combination with feature selection, anomaly detection, and clustering analysis, can be used enhancing intrusion detection's precision. By pinpointing significant decreased input characteristics and developing robust models, network security can be enhanced, and potential threats can be mitigated.

2. Literature Survey

The need to combine an enterprise's operational data and to give centralised, hence regulated access to that data is one of the primary drivers behind the usage of database systems. On the other side, computer network technology encourages a way of working that opposes all attempts at centralization. It may first seem difficult to comprehend how these two opposing strategies might be combined to create a technology that is more potent and promising than each one used alone. It's crucial to understand that none of these phrases necessitates the other. Integration may be accomplished without centralised control, and distributed database technology aims to do just that. Within this context, distributed database systems ought to be seen as tools that could facilitate and improve distributed processing. It is acceptable to compare what database technology has previously offered to what distributed databases could deliver to the data processing industry. Without a doubt, the creation of general-purpose, flexible, effective distributed database systems has tremendously facilitated the process of creating distributed software [1].

The issues with testing global consistency progressively in response to modifications made to the base relations without having access to all of these base relations are the most crucial elements in this article. Total data availability cannot be assumed in many application areas, and some data may be so expensive that using them should only be done as a last resort. Finding

ISSN: 2326-9865
DOI: https://doi.org/10.17762/msea.v70i1.2300

tests that are the most generic (which we refer to as Complete Local Tests) and that are quick to produce and run is a problem for consistency maintenance. We offer comprehensive local tests under additions and deletions to the local relations for restrictions where the predicates for the distant relations do not appear more than once. As safe, non-recursive Datalog queries on the local relations, these tests may be described. These findings also hold true for other negation-based restrictions that are not conjunctive [2].

Today's natural sciences analyse vast amounts of diverse data, which has resulted in the emergence of several independent, highly specialised databases with complex linkages. We offer exploratory queries as a user-friendly approach for correlating this complicated, increasing network of sources. Exploratory searches don't require much source network expertise because they are loosely organised. Evaluation of several dispersed queries often occurs when evaluating an exploratory query. By putting forth a number of multi-query optimisation methods that calculate a global assessment plan while minimising the overall communication cost, a significant bottleneck in distributed settings, we tackle the optimisation issue associated with exploratory inquiries. Such scattered searches have the potential to swiftly increase in quantity. Given that it is NP-hard to calculate an ideal global assessment plan, the suggested methods must be heuristics. Finally, we offer an implementation of our algorithms together with tests that highlight their potential for both the multi-query optimisation of sizable batches of ordinary queries as well as the optimisation of exploratory inquiries [3].

In reaction to modifications (insertions, deletions, or updates) to the relations, the author presents incremental evaluation techniques to compute changes to materialised views in relational and deductive database systems. The view definitions may include UNION, negation, aggregation (such as SUM, MIN), linear recursion, and general recursion. They may be written in SQL or Datalog. We start by introducing a counting technique that keeps track of how many different derivations (counts) there are for each derived tuple in a view. The technique is compatible with duplicate and set semantics. We outline the technique for non-recursive views (with negation and aggregation) and demonstrate that it is possible to determine a tuple's count with little to no additional cost above the cost of obtaining the tuple itself. Because it calculates precisely the view tuples that are added or removed, the algorithm is ideal. Notably, we only keep track of the quantity of derivations rather than the actual derivations. Next, we introduce the Delete and Rederive algorithm (DRed), which allows for negation and aggregation while still allowing for incremental maintenance of recursive views. A subset of the tuples that need to be destroyed is first eliminated using the algorithm, and some of those tuples are then rederivated. When the view definition itself is changed, the algorithm can still be applied [4].

When the time required by an incremental algorithm to conduct an update can be expressed as a function of the sum of the sizes of the changes to input and output, the method is said to be limited. If a dynamic issue lacks a bounded incremental method inside a computing model, it is said to be unbounded with regard to that model. New lower-bound and upper-bound results are presented in the paper with regard to a group of algorithms known as locally persistent algorithms. In classifying dynamic issues according to their incremental complexity with regard to locally persistent algorithms, our results and some previously known ones help to

clarify how the complexity hierarchy is organised. These findings distinguish between the groups of polynomial problems that are bounded, naturally exponentially bounded, and unbounded. Asymptotic worst-case analysis and expressing the computation's cost as a function of input size are two methods frequently used to gauge an algorithm's temporal complexity. However, this type of analysis is not always very useful for incremental algorithms. The term "incremental algorithm" refers to a method for a dynamic issue. Numerous incremental algorithms have been proposed, but when the computation's cost is formulated as a function of the size of the (current) input, they run in worst-case asymptotically no faster than starting from scratch [5].

3. Proposed System

As a result, the current intrusion detection methods have concentrated on the problems of feature selection or dimensionality reduction. It favours the collection of characteristics that have low intracorrelation but strong correlation with the class. The IG assesses characteristics by calculating their information gain relative to the class. It initially uses an MDL-based discretization approach to discretize numerical properties. Using a correlation-based feature selection technique, the original high dimensional database's useless data is eliminated. A technique for intrusion detection that addresses the issues of uncertainty brought on by incomplete and unclear information. The resulting dataset is then utilised for the training and testing of the classifier, and the procedure is repeated until the feature-based classifier outperforms the original dataset in terms of the pertinent performance criteria.

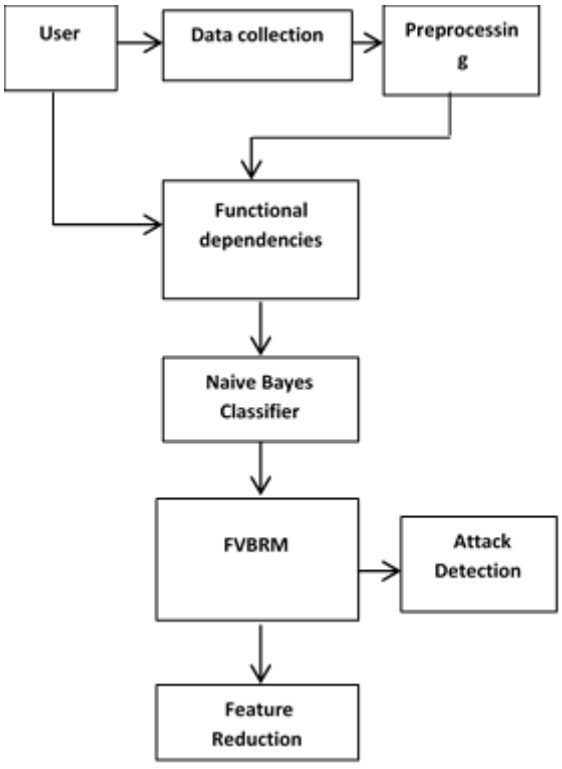

Fig 1: Data Flow Diagram

Three primary performance metrics—classification accuracy, TPR, and FPR of the system—are taken into account to assess the usefulness of a feature. By starting with the set of all features and removing each feature one at a time until the classifier's accuracy fell below a certain level, we were able to determine the crucial set of features. The most recognised text on data mining and machine learning, now in its eagerly awaited third edition, will teach you all you need to know about setting up inputs, analysing outputs, assessing outcomes, and the algorithmic techniques at the core of effective data mining. Comprehensive updates take into account the technical advancements and modernizations that have occurred in the field since the publication of the previous edition. These updates include new information on Massive data sets, multi-instance learning, ensemble learning, data transformations, plus an updated rendition of the well-known Weka machine learning algorithm are all included tool. On a dataset with features, we apply the naive bayes classifier and examine its performance indicators, such as accuracy. Next, we are determining the threshold values. The following are a few of benefits of the suggested approach:

- It raises the level of precision.
- Superior performance.
- Increasing classification precision.
- It is more flexible and efficient.
- Simple threshold values to locate

The next section provides an explanation of the many phases that are involved in putting the suggested technique into practise:

1. Dataset Collection

A data set is often the contents of a single database table or statistical data matrix, where each row denotes a particular component of the data set in question while each column denotes a particular variable. The data set contains values for each variable, including an object's height and weight, for each data set participant. A datum is a name for any value. According on the number of rows, the data set may include information for one or more members. 40 tables of fields make up the KDD data collection. Dataset for data analysis and clustering. It consists of many clusters of URL data. The information that displays the group of fields that the dataset query on the data source will return.

2. Preprocessing

This procedure has made use of the KDD dataset. The raw data must first be loaded before being preprocessed. The tasks involved in data preparation and filtering might take a long time to process. Data analysis that has not been thoroughly checked for these issues may yield false findings. Therefore, before performing an analysis, it is crucial to consider the representation and quality of the data.

3. Naive Bayes Classifier

This talk will concentrate on

ISSN: 2326-9865
DOI: https://doi.org/10.17762/msea.v70i1.2300

Methods for learning and categorization based on probability theory. A crucial component of probabilistic learning and classification is the Bayes theorem. uses the prior probability for each category in the absence of any item-specific information.

Given a description of an item, categorization generates a posterior probability distribution across all potential categories.

4. DOS Attack Detection

- Putting items on the process table
  – Achieved by recursive forking
  – Stops additional users from starting new processes
- Adding files to the system
  – By regularly writing a large amount of data to the file system
  – Stops other users from making changes to files
  – May lead to a system crash
- Sending traffic outward that clogs the network link
  – Through the use of a programme that continuously sends phoney network traffic
  – Consumes network traffic and CPU resources.
- Assault on land
  – Sends a fake packet to a target with the identical source and target IP and port numbers, taking down the network services of the victimised target.
  – Multiple Land attack packets are sent to various ports.

5. Feature Reduction

- Filtering strategy
  – It assigns a value to each feature or feature subset without regard to the predictor (classifier).
  – Applying univariate techniques think about one thing at a time.
  – Consider several variables at once while using multivariate approaches.
- Wrapper strategy
  – Evaluates (many) characteristics or feature subsets using a classifier.
- Embedding strategy
  – Builds a (single) model using a subset of internally chosen features and a classifier.
  – Train the predictor using training data for each feature subset.
  – The feature subset that performs best on validation data should be chosen.
  – If you wish to lower the variance (cross-validation), repeat and average.
  – Testing using test data.

4. RESULTS

In order to build an IDS that is successful and efficient computationally, this paper investigates the effectiveness of three popular feature selection strategies. To find attacks on the four attack types—Probe, DoS, U2R, and R2L—the NSL KDD dataset is employed. Anomaly detection is used to identify out-of-the-ordinary behaviour. A Wrapper evaluates the value of features using the intended learning method itself, whereas a filter does so using heuristics based on the

ISSN: 2326-9865
DOI: https://doi.org/10.17762/msea.v70i1.2300

general aspects of the data. These two techniques are widely employed for feature reduction. The effectiveness of a feature is evaluated using three key performance measures.

The most authoritative book on data mining and machine learning, now in its widely anticipated third edition, will teach you all you need to know about configuring inputs, analysing outputs, evaluating outcomes, and the mathematical approaches at the heart of efficient data mining. We use the naïve bayes classifier and look at its performance metrics, such as accuracy, on a dataset containing features. Next, we're figuring out the threshold values that are displayed in the screenshots below.

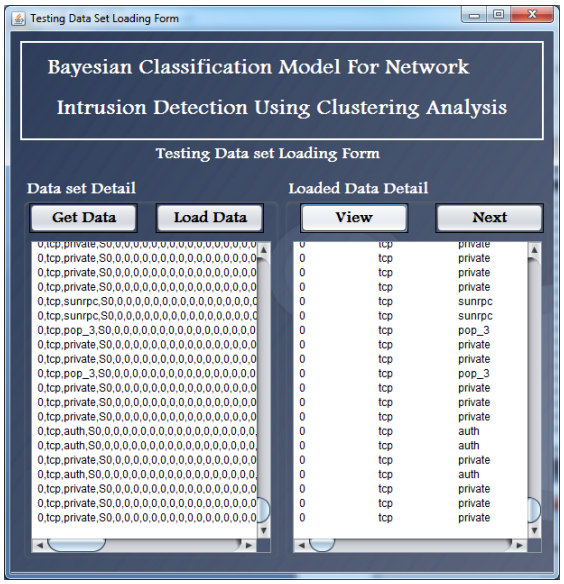

Fig 2: Loaded Data Details

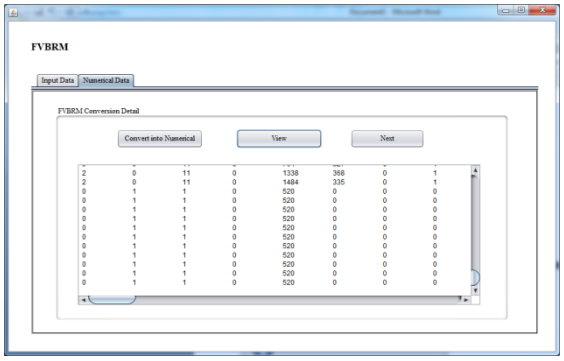

Fig 3: FVBRM Conversion Details

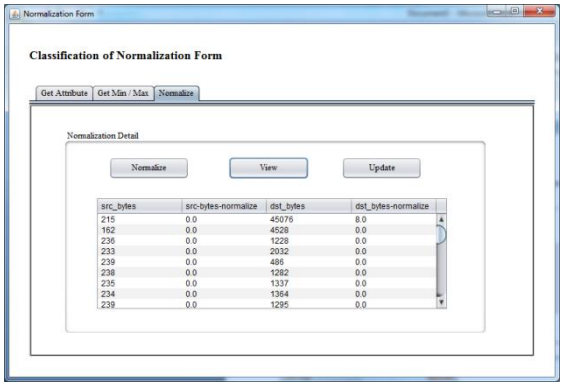

Fig 4: Classification of Normalization Form

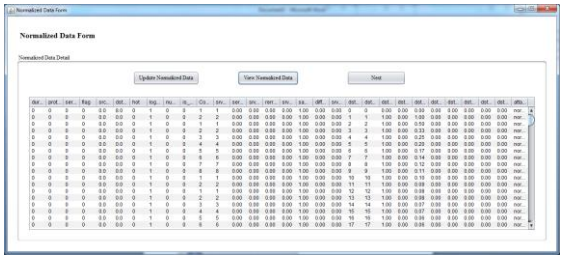

Fig 5: Normalized Data

5. Conclusion

The Bayesian Classification Model, a probabilistic classification method, has gained significant attention for its ability to accurately classify network traffic into normal and anomalous behavior. It is particularly suitable for Network Intrusion Detection (NID), where the goal is to identify potential threats in real-time. Clustering Analysis, a widely used unsupervised learning technique, can be utilized to group similar network traffic data together, reducing the overall computational burden of the Bayesian Classification Model. The Bayesian Classification Model offers several advantages over traditional rule-based and signature-based approaches, such as adaptability to changing network conditions and detection of unknown threats. It can also be further improved by incorporating domain knowledge and tuning hyper parameters. The Bayesian Classification Model is a powerful tool for NID, with its ability to adapt to changing network conditions and detect unknown threats, coupled with its probabilistic reasoning, making it an attractive option for cyber security professionals.

Reference

1. R. Kohavi, G.H. John, Wrappers for feature subset selection, Artificial Intelligence, 1(2) (1997) 273–324.
   Bace, R. (2000). Intrusion Detection. Macmillan Technical Publishing.
2. Markou, M. and Singh, S., Novelty Detection: A review, Part 1: Statistical Approaches, Signal Processing, 8(12), 2003, pp. 2481-2497.
3. Liu H ,Setiono R, Motoda H, Zhao Z Feature Selection: An Ever Evolving Frontier in Data Mining, JMLR: Workshop and Conference Proceedings 10: 4-13 The Fourth Workshop on Feature Selection in Data Mining.

4. I.H.Witten, E.Frank, M.A. Hall " Data Mining Practical Machine Leanrning Tools & Techniques" Third edition, Pub. – Morgan kouffman.
5. Mark A. Hall, Correlation-based Feature Selection for Machine Learning, Dept of Computer Science, University of Waikato.http://www.cs.waikato.ac.nz/~mhall/thesis.pdf.
6. j.Han ,M Kamber, Data mining : Concepts and Techniques. San Francisco, Morgan Kauffmann Publishers(2001).
7. Wafa' S.Al-Sharafat, and Reyadh Naoum "Development of Genetic-based Machine Learning for Network Intrusion Detection" World Academy of Science, Engineering and Technology 55, 2009
8. Ms.Nivedita Naidu, Dr.R.V.Dharaskar "An effective approach to network intrusion detection system using genetic algorithm", International Journal of Computer Applications (0975 – 8887) Volume 1 – No. 2, 2010.
9. James P. Anderson. Computer Security Threat Monitoring and Surveillance, 1980. Last accessed: November 30, 2008. http://csrc.nist.gov/publications/history/ande80.pdf.
10. Dorothy E. Denning. An Intrusion-Detection Model. IEEE Transactions on Software Engineering,13(2):222–232, 1987. IEEE.