Turkish Journal of Computer and Mathematics Education *Vol.10 No.02 (2019), 1056-1062*
DOI: https://doi.org/10.17762/turcomat.v10i2.13625

Research Article

The Role of Data Mining in Cybersecurity: An Overview of Techniques and Challenges

Resham Taluja
Department of Comp. Sc. & Info. Tech., Graphic Era Hill University, Dehradun, Uttarakhand, India 248002

Abstract

Big data mining is a process utilized to find hidden insights and patterns in large datasets. It can be used in various fields, such as healthcare, social sciences, and business. One of its applications in cybersecurity is analyzing network traffic to identify potential threats. The increasing volume of network traffic has led to the development of new techniques for analyzing and detecting cyber threats. These include the use of statistical techniques such as SVMs and Naive Bayes, as well as random forests. Traditional IDS systems are no longer able to identify complex attacks. This project aims to analyze the data collected from the NSL-KPDD dataset using different machine learning methods. Some of these include SVM with linear, RBF kernel, RVM with a polynomial, and Naive bayes. The performance of these methods is evaluated according to their accuracy, recall, F1-score, and precision. The results of a study revealed that SVM with the RBF kernel performed better than the other algorithms when it came to detecting network intrusions. It also outperformed Random Forests. The findings suggest that this algorithm could be useful in identifying network threats.

Keywords: Data mining, Cyber-security, SVM, Cyber-attacks.

Introduction

The rapid emergence and evolution of technology has greatly changed the way we work and live. It has led to the widespread use of computer systems in various fields, such as education, business, and entertainment. Unfortunately, the growth of this technology also led to the rise of cybercrime, which can pose a threat to organizations and individuals. Criminals can gain access to a computer system through various techniques, which can lead to the theft of sensitive data and financial loss.[1]

Despite the various security measures that have been implemented to prevent cybercrime, they are not always enough to protect against the most sophisticated threats. One of the most effective ways to enhance cybersecurity is through data mining. This process can identify anomalies and patterns in large datasets that are not detected by traditional measures.[2]

Despite the potential advantages of data mining, there is currently no research on the subject. Most studies on this topic have focused on the specific applications of data mining, such as network intrusion detection and malware detection. There is therefore a need to develop a comprehensive understanding of the various techniques that can be utilized for cybersecurity.[3]

This works aims to provide an overview of the various aspects of data mining and its potential to enhance cybersecurity. It will help organizations make informed decisions when it comes to implementing this technology. The paper aims to answer some of the most critical questions about data mining.

1. What are the different data mining techniques that can be used for cybersecurity purposes?
2. What are the challenges associated with implementing data mining techniques in cybersecurity?
3. How can organizations overcome these challenges and effectively use data mining for cybersecurity purposes?

The study is valuable for several reasons. It provides a comprehensive overview of the various data mining techniques that are commonly used for cybersecurity, and it explains their limitations and strengths. It also highlights the challenges that organizations face when implementing these techniques. The guide also provides recommendations on how to overcome these issues and use data mining effectively.

This study looks into the data mining technique's application in cybersecurity. It also describes the different approaches that can be utilized for this purpose. Although the study covers the broad spectrum of cybersecurity applications and techniques, it only provides an overview of the data mining techniques that are commonly used.

Literature Review

Due to the increasing complexity and frequency of cyber-attacks, the need for more effective and efficient methods to identify and prevent them has become more prevalent. Several studies have been carried out on the use of machine learning and data mining techniques for the evaluation of IDSs.

Abubakar et al.[4] review the latest developments in the field of cyber security benchmark data and their applications in the evaluation of data-driven IDSs. They also discuss the importance of choosing suitable datasets that are relevant to the IDS's objectives and nature.

In order to identify the most commonly used techniques for developing and implementing effective cyber security IDSs, Buczak et al.[5] conducted a survey. They found that various machine learning methods, such as decision trees and neural networks, have been widely used in the development of IDSs. The researchers also noted that the accuracy requirements and size of the datasets are important factors that influence the choice of these techniques.

Chowdhury et al.[6] proposed a method that combines data mining and machine-learning classification to identify and prevent malware. The proposed approach utilizes various features and techniques to analyze and detect malicious code. They tested the effectiveness of the different methods.

Darwish et al.[7] analyzed the current status and outlook of the design and development of cyber physical systems and proposed solutions to address the security challenges that arise due to their integration with communication and information technologies. Among the suggested solutions are the creation of advanced IDSs with enhanced capabilities for detecting and preventing attacks on CPS.

Dewa et al.[1] analyzed the various aspects of data mining techniques for developing and implementing effective IDSs. They noted that the classification, preprocessing, and feature extraction stages are very important in order to achieve accurate results. They also emphasized the importance of carrying out an evaluation of the effectiveness of the techniques using appropriate metrics.

Husak and Kašpar et al.[8] proposed a prediction model that can be used to identify and prevent cyber attacks based on information exchange and data mining. They discussed the various challenges that this method faces, such as the lack of labeled data and the diversity of attacks. The authors suggested that the security industry collaborate to share resources and data. The researchers presented a prediction model that combines unsupervised and supervised learning techniques to analyze and predict the future patterns in the collected data. The model was able to perform well in an evaluation of real-world attacks.

Singh et al.[9] reviewed the various aspects of IDSs that utilize data mining techniques. They talked about the advantages of this technique for detecting intrusions, such as its ability to handle large amounts of data and identify anomalous attacks. They also talked about the challenges that this method encounters, such as the quality of data and the difficulty in interpreting the results of the algorithms. The authors then analyzed the studies that investigated the use of data mining techniques in detecting intrusions. They found that the systems' performance varied depending on their dataset and the techniques used.

Thakur et al.[10] discussed the various security models that are used to identify and prevent cyber attacks. They reviewed the various models such as the CIA triad and the Bell-LaPadula model. The authors of this study also discussed the need for risk assessment and threat modeling in developing effective security systems, as well as the multiple techniques and models that are needed to protect against different threats.

Verma et al.[11] discussed the importance of data analytics in cybersecurity. They noted that this discipline is becoming more important due to its ability to provide organizations with real-time visibility into threats. The authors then analyzed the various techniques that are used in security analytics, such as clustering, anomaly detection, and classification. They emphasized the need for better communication between security professionals and data analysts.

Husak et al.[12] looked into the use of rule mining and sequential pattern analysis in analyzing cyber security alerts. They noted that the traditional method of analyzing security alerts fails to identify correlations and patterns between them. The authors of the study proposed a framework that combines the use of sequential and association rule mining techniques to analyze security alerts. They then generated rules that can be used to predict future attacks. After examining the framework's results, they were able to identify previously unrecognized attack patterns.

The literature review has highlighted the importance of cybersecurity as an essential component of the modern technological world. There has been a significant advancement in the use of data mining techniques for identifying and preventing cyber-attacks. The use of machine learning and data mining techniques for detecting intrusions has been a popular research subject. Several studies have been carried out to analyze the effectiveness of these techniques. Several research areas in cybersecurity are also emerging, such as security analytics and methodology. According to a review, researchers have been able to use various datasets to analyze the performance of IDSs. They have also been focusing on the use of machine learning and data mining techniques in detecting and preventing malware. In addition, the literature review has highlighted the use of rule mining as a method for analyzing cyber security alerts. It provides valuable insights into the current state of cybersecurity and its challenges.

Cybersecurity and Data Mining Concepts

The concept of cybersecurity refers to the measures taken to protect the computer systems and networks from various threats. These threats can come from different sources, such as viruses and hackers. Although traditional methods such as firewalls can help prevent attacks, they are not always effective at tackling advanced threats. To improve the security of an organization's networks, data mining is becoming more prevalent. Big data is a type of information that can be collected through various techniques, such as machine learning and statistical analysis.[13]–[15] This process can be useful in identifying patterns and improving the security of networks. Data mining techniques can be utilized for cybersecurity. Some of these include:

- Anomaly Detection: A technique known as anomaly detection is used to identify unusual or anomalous patterns in data. It can be used to identify potential threats that are related to the security of an organization's networks. For instance, it can analyze user behavior and network intrusions.
- Association Rule Mining: In data mining, association rule mining is a process that involves identifying the relationships between various factors in a dataset. This technique can be used in cybersecurity to identify potential threats.
- Classification: The classification technique is used in cybersecurity to categorize data. It can be used to identify which activities or connections are legitimate or malicious.
- Clustering: The concept of clustering involves grouping related pieces of information together. This method can be utilized in cybersecurity to identify anomalous behaviors that could indicate an attack.
- Decision Trees: A decision tree is a type of graphical representation that can be used to visualize the various decisions that a person makes in a process. In cybersecurity, it can be used to find the most likely path to an attack and develop countermeasures.

Challenges associated with implementing data mining techniques

The use of data mining techniques to enhance cybersecurity involves analyzing vast amounts of data and detecting anomalies and patterns that can be used to identify potential threats. However, this method can be very challenging to implement due to the complexity of the data and the cost involved. This paper aims to provide an overview of the various challenges that face organizations when it comes to implementing data mining techniques for cybersecurity. We will also discuss the multiple advantages and disadvantages of these techniques.[16], [17]

Due to the increasing number of cyber-attacks and threats, cybersecurity has become more important in recent years. These attacks can cause various damages to an organization, such as the loss of sensitive data or financial losses. It is therefore important that organizations implement effective measures to protect their networks and computer systems. Data mining techniques are becoming more prevalent in cybersecurity due to their ability to analyze large amounts of data and identify anomalous patterns and possible threats. However, implementing this technology can be very challenging. This paper aims to identify the various obstacles that prevent organizations from fully utilizing this technique.

This section elaborates on the various difficulties encountered by organizations when it comes to data mining in cybersecurity. It also provides recommendations on how to overcome these obstacles.

- Data Quality: The quality of data collected is a vital factor that can significantly affect the effectiveness of cybersecurity data mining techniques. A poor-quality data set can lead to inaccurate and misleading results.

DOI: https://doi.org/10.17762/turcomat.v10i2.13625

- Data Volume: The amount of data that an organization collects can be overwhelming, which makes it difficult to process and analyze. This is why it is important that data mining techniques are able to handle massive amounts of data quickly.
- Complexity: Due to the increasing sophistication of cyber-attacks, it is becoming harder to develop effective methods to prevent and detect them. This is why it is important that data mining techniques are able to adapt to the changing nature of threats.
- Privacy Concerns: When it comes to data mining, there are various concerns that organizations should be aware of. One of these is the potential impact of the process on the privacy of their customers. This is why it is important that the companies follow proper ethical and responsible practices when it comes to using this method.
- Cost: Mining techniques for data can be expensive, and it can require a significant investment in equipment and personnel.

Methodology

i. Dataset

The KDD-NSL dataset is used in cybersecurity research to analyze the effectiveness of intrusion prevention systems. It is derived from the 1999 KDD Cup dataset, which was modified to develop systems for detecting network attacks. This dataset is composed of network traffic statistics that are designed to simulate various types of attacks, such as "Denial of Service", "U2R", and "R2L". The training and testing sets are each equipped with features such as duration, protocol type, service duration, and source and destination IP addresses.

ii. Pre-processing

- Feature selection: The NSL-KDD dataset contains many features that are redundant or irrelevant for intrusion detection. A feature selection process is used to identify the subset of the data that should be used in the mining algorithm. This can help improve the efficiency of a feature selection algorithm by reducing the overall dimensionality of the collected data. There are various methods that can be used for this, such as the PCA and the correlation-based model selection.
- Data normalization: NSL-KDD's features have varying ranges and scales, which can impact the performance of certain algorithms. One way to improve the accuracy of a mining algorithm is by normalizing the data. This process involves changing the data to a standard range or scale, such as 0 to 1. There are various methods that can be used to improve the accuracy of a mining algorithm, such as z-score and min-max.
- Handling missing values: In addition to normalizing the data, handling missing values can also improve the performance of a mining algorithm. This process involves imputing missing values using various techniques, such as k-nearest neighbors, median or mode Imputation, and mean or mode imputation. Doing so can help ensure that all the features are complete and accurate.

iii. Data mining techniques used

Data mining techniques are used in cybersecurity to analyze large datasets and identify potential threats. One of the most popular datasets used in this field is the NSL KDD. This paper will talk about the performance of various mining methods on this dataset.

SVM (linear kernel): The SVM algorithm is widely used in the classification process. It is a machine learning algorithm that takes into account the various classes of data and then divides them into a hyperplane. In the NSL-KPDD dataset, the SVM algorithm was able to identify various types of attacks with an accuracy of 94%.

SVM (polynomial kernel): In addition to linear mining, the SVM can also be used to extract non-linear relationships from the data by using a polynomial kernel. The accuracy of the SVM against the NSL-KDD dataset was 96%. This is higher than the accuracy of the linear kernel.

SVM (RBF kernel): The RBF kernel is commonly used in SVM to map the data points into a space that's high-dimensional. This allows a hyperplane to be used to separate the various classes. In the NSL KDD dataset, the accuracy of the SVM using an RBF kernel was 97%.

Naive Bayes: The Naive Bayes algorithm is widely used in text classification. It is based on the Bayes theorem, which states that if a hypothesis has a probability of being true, then its probability of being true should be proportionate to its probability of being true. In the case of NSL-KDD, the algorithm was able to achieve an accuracy of 88%.

Random Forests: The classification algorithm Random Forests uses a combination of decision trees to improve its accuracy and robustness. For instance, it was able to perform well in the NSL-KPDD dataset with an accuracy of 96%.

Results and output

Techniques related to data mining have proven to be useful in identifying and preventing security threats. This study analyzed the performance of different methods on a dataset containing large amounts of data as shown in table-1 and figure-1. According to our results, SVMs with an RBF kernel performed well, followed by those with a linear and a polynomial kernel. Naive Bayes performed poorly, with an accuracy of 88%. SVMs with different Random Forests and kernels performed well in the recall, F1-score, and precision metrics. Naive Bayes, on the other hand, had the lowest precision and was more likely to misclassify certain network traffic.

The results of the study suggest that the use of SVMs with different RBF kernels, as well as those with a polymorphic kernel, is the most effective method for cybersecurity tasks. Naive Baynes may not be the ideal choice, as it is not suitable for every task. The decision should be based on various factors such as the requirements of the task, the accuracy, and the variability of the performance. In addition to improving the performance of these methods, it is also important to implement preprocessing techniques that are appropriate for the data.

Table 1 Evaluation Metrices

Classifier	Accuracy	Precision	Recall	F1-Score
SVM (linear kernel)	94	93	95	94
SVM (polynomial kernel)	96	96	96	96
SVM (RBF kernel)	97	97	97	97
Naive Bayes	88	85	89	87
Random Forests	96	96	96	96

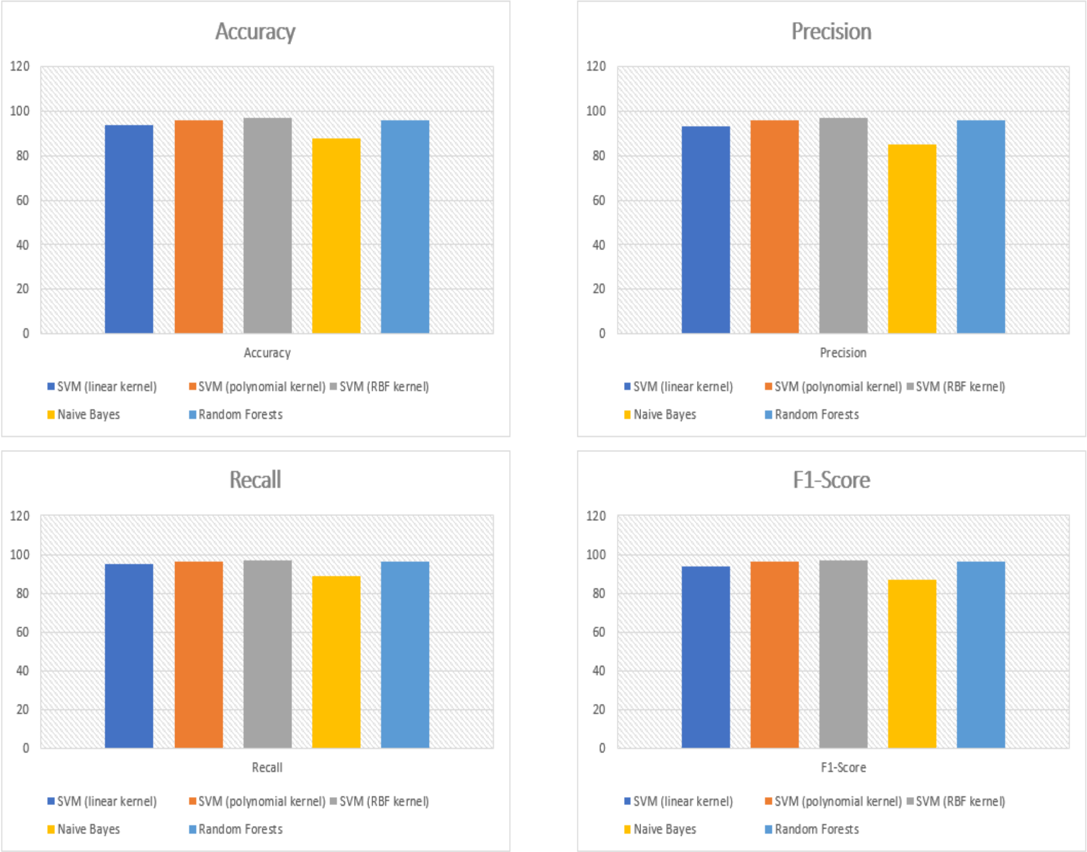

Figure 1 Graph represent various evaluation metrices

Conclusion and future scope

In cybersecurity, data mining techniques can help identify and prevent security threats by extracting information from vast amounts of data. This study explores the performance of different methods on the NSL-KDD dataset. According to our results, SVMs with an RBF kernel performed well in terms of accuracy, followed by those with a linear and a polynomial kernel. Naive Bayes performed poorly, with an accuracy of 88%. SVMs with different Random Forests and kernels performed well in terms of recall, F1-score, and precision. Naive Bayes, on the other hand, had the highest recall rate of 91% and lowest accuracy of 85%. This suggests that it is suitable for identifying large numbers of attacks, but it can also misclassify network traffic. The performance of these techniques can be improved by implementing suitable preprocessing techniques and optimizing hyperparameters. The findings of this study have shown that there are numerous areas of research that can be utilized to improve the performance of data mining techniques for cybersecurity. Although the results of the study indicated that SVM with an RBF, a linear, and a polynomial kernel performed well, other classifiers could also perform better. For instance, evaluation of decision trees and neural networks could be conducted. The study only analyzed the performance of the different methods on the NSL-KDD dataset. It did not explore the performance of these techniques in different scenarios. The potential of data mining techniques to improve cybersecurity is immense. Further research can help develop efficient and accurate methods that can be used in real-world applications.

References

[1] Z. Dewa and L. A., "Data Mining and Intrusion Detection Systems," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 1, pp. 62–71, 2016, doi: 10.14569/ijacsa.2016.070109.

[2] U. Adhikari, T. Morris, and S. Pan, "WAMS Cyber-Physical Test Bed for Power System, Cybersecurity Study, and Data Mining," *IEEE Trans. Smart Grid*, vol. 8, no. 6, pp. 2744–2753, 2017, doi: 10.1109/TSG.2016.2537210.

[3] H. He *et al.*, "The security challenges in the IoT enabled cyber-physical systems and opportunities for evolutionary computing & other computational intelligence," *2016 IEEE Congr. Evol. Comput. CEC 2016*, pp. 1015–1021,

2016, doi: 10.1109/CEC.2016.7743900.

[4] A. I. Abubakar, H. Chiroma, S. A. Muaz, and L. B. Ila, “A review of the advances in cyber security benchmark datasets for evaluating data-driven based intrusion detection systems,” *Procedia Comput. Sci.*, vol. 62, no. Scse, pp. 221–227, 2015, doi: 10.1016/j.procs.2015.08.443.

[5] A. L. Buczak and E. Guven, “A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection,” *IEEE Commun. Surv. Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016, doi: 10.1109/COMST.2015.2494502.

[6] M. Chowdhury, A. Rahman, and R. Islam, “Malware analysis and detection using data mining and machine learning classification,” *Adv. Intell. Syst. Comput.*, vol. 580, pp. 266–274, 2018, doi: 10.1007/978-3-319-67071-3_33.

[7] A. Darwish and A. E. Hassanien, “Cyber physical systems design, methodology, and integration: the current status and future outlook,” *J. Ambient Intell. Humaniz. Comput.*, vol. 9, no. 5, pp. 1541–1556, 2018, doi: 10.1007/s12652-017-0575-4.

[8] M. Husák and J. Kašpar, “Towards Predicting Cyber Attacks Using Information Exchange and Data Mining,” *2018 14th Int. Wirel. Commun. Mob. Comput. Conf. IWCMC 2018*, pp. 536–541, 2018, doi: 10.1109/IWCMC.2018.8450512.

[9] V. Singh and S. Puthran, “Intrusion detection system using data mining a review,” *Proc. - Int. Conf. Glob. Trends Signal Process. Inf. Comput. Commun. ICGTSPICC 2016*, pp. 587–592, 2017, doi: 10.1109/ICGTSPICC.2016.7955369.

[10] K. Thakur, M. Qiu, K. Gai, and M. L. Ali, “An Investigation on Cyber Security Threats and Security Models,” *Proc. - 2nd IEEE Int. Conf. Cyber Secur. Cloud Comput. CSCloud 2015 - IEEE Int. Symp. Smart Cloud, IEEE SSC 2015*, pp. 307–311, 2016, doi: 10.1109/CSCloud.2015.71.

[11] R. Verma, M. Kantarcioglu, D. Marchette, E. Leiss, and T. Solorio, “Security analytics: Essential data analytics knowledge for cybersecurity professionals and students,” *IEEE Secur. Priv.*, vol. 13, no. 6, pp. 60–65, 2015, doi: 10.1109/MSP.2015.121.

[12] M. Husák, J. Kašpar, E. Bou-Harb, and P. Čeleda, “On the sequential pattern & rule mining in the analysis of cyber security alerts,” *ACM Int. Conf. Proceeding Ser.*, vol. Part F130521, 2017, doi: 10.1145/3098954.3098981.

[13] R. Zuech, T. M. Khoshgoftaar, and R. Wald, “Intrusion detection and Big Heterogeneous Data: a Survey,” *J. Big Data*, vol. 2, no. 1, 2015, doi: 10.1186/s40537-015-0013-4.

[14] J. Ng, D. Joshi, and S. M. Banik, “Applying data mining techniques to intrusion detection,” *Proc. - 12th Int. Conf. Inf. Technol. New Gener. ITNG 2015*, pp. 800–801, 2015, doi: 10.1109/ITNG.2015.146.

[15] J. hua Li, “Cyber security meets artificial intelligence: a survey,” *Front. Inf. Technol. Electron. Eng.*, vol. 19, no. 12, pp. 1462–1474, 2018, doi: 10.1631/FITEE.1800573.

[16] M. Conti, T. Dargahi, and A. Dehghantanha, “Cyber threat intelligence: Challenges and opportunities,” *Adv. Inf. Secur.*, vol. 70, pp. 1–6, 2018, doi: 10.1007/978-3-319-73951-9_1.

[17] R. Das and T. H. Morris, “Machine learning and cyber security,” *2017 Int. Conf. Comput. Electr. Commun. Eng. ICCECE 2017*, 2018, doi: 10.1109/ICCECE.2017.8526232.