Automation and Machine Learning (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2024.050105
ISSN 2516-5003 Vol. 5 Num. 1

Impact of High-Dimensional Feature Selection Strategies Based on Machine Learning on Malicious Document Detection

Jiaoli Zhou

Hainan College of Science and Technology, Haikou, 571126, China

Keywords: Machine Learning, High-Dimensional Feature Selection, Malicious Document Detection, Performance Evaluation, Model Complexity

Abstract: This paper explores the impact of high-dimensional feature selection strategies based on machine learning in malicious document detection. Addressing the existing issues in the field of malicious document detection, a new strategy for high-dimensional feature selection utilizing machine learning techniques is proposed. Through empirical research on multiple datasets, the effectiveness of this strategy in enhancing the performance of malicious document detection is evaluated. The results show that high-dimensional feature selection can maintain high accuracy while reducing model complexity and improving detection efficiency.

1. Introduction

With the widespread use of the internet, the number and complexity of malicious documents have been increasing, posing serious challenges to information security. In the field of malicious document detection, traditional feature extraction methods often lead to a high-dimensional feature space, creating difficulties in model training and prediction. To address this issue, this paper proposes a high-dimensional feature selection strategy based on machine learning. By automatically selecting the most relevant features during the training process, we aim to improve the performance of malicious document detection models and reduce computational costs.

2. Analysis of the Current Status of Malicious Document Detection

2.1. Definition and Classification of Malicious Documents

As a significant research object in the field of cybersecurity, the definition of malicious documents encompasses various forms including viruses, trojans, and worms. In this section, we first clarify the broad definition of malicious documents, which includes any file containing executable malicious code. Further subdividing, we classify malicious documents considering their specific threat types and attack methods.

From the perspective of threat types, malicious documents can be divided into two main categories: malware and malicious scripts. Malware typically exists in the form of executable files, including viruses, worms, trojans, etc., while malicious scripts focus more on using scripting languages like

JavaScript, VBScript, etc., to lurk in systems through web pages or documents.[1-2]

Moreover, we have classified malicious documents based on attack methods, including social engineering attacks, exploitation of vulnerabilities, phishing, etc. This detailed classification helps us better understand the diversity and complexity of malicious documents, providing more targeted directions for subsequent detection methods.

Understanding these different forms of malicious documents is crucial for effective detection. Attackers constantly change the form of malicious documents in an attempt to evade traditional detection methods. Therefore, we need to delve deeper into the characteristics of malicious documents to better adapt to the evolving cybersecurity threats.[3]

2.2. Existing Methods for Malicious Document Detection

2.2.1. Traditional Rule and Signature-Based Methods

Traditional methods for detecting malicious documents primarily rely on rules and signatures of known threats. The core idea of this method is to manually formulate a series of rules and signatures by security experts to identify known viruses, worms, and other malicious documents. These rules and signatures act as templates, enabling the system to quickly match known threats and take appropriate defense measures. However, this method is insufficient against unknown threats, as rules and signatures cannot cover all possible malicious variants, making them easy targets for attackers to bypass, leading to detection failures.

2.2.2. Machine Learning-Based Methods

With the rapid development of machine learning technology, an increasing number of studies are applying it to malicious document detection. These methods train models to learn the features of malicious documents, achieving adaptive detection of unknown threats. Compared to traditional methods, machine learning-based malicious document detection has better generalization ability and can capture potential unknown threats. However, due to the high-dimensional feature space of malicious documents, traditional machine learning models face the challenges of the curse of dimensionality and computational complexity when dealing with high-dimensional data.[4]

2.2.3. Challenges in High-Dimensional Feature Space

In high-dimensional feature space, malicious document detection faces a series of challenges. First, the curse of dimensionality leads to a decrease in the model's generalization ability, making it prone to overfit known threats and unable to adapt to new, unknown threats. Second, the computational complexity increases significantly, making the training and prediction process of the model time-consuming and resource-intensive, thereby affecting the system's real-time performance and efficiency. To overcome these issues, an efficient high-dimensional feature selection strategy is urgently needed, which means reducing the dimensionality of the feature space while maintaining key information, to enhance the model's performance and adaptability. The introduction of this strategy is expected to solve the challenges faced in high-dimensional data for malicious document detection, providing a more reliable and efficient defense mechanism for the system.[5]

3. High-Dimensional Feature Selection Strategies Based on Machine Learning

3.1. Overview of High-Dimensional Feature Selection

In the field of malicious document detection, the high-dimensional feature space presents

challenges for model training and performance. This chapter aims to find an effective solution to this problem, namely high-dimensional feature selection. We will explore three main feature selection techniques: filtering, wrapping, and embedding.

3.1.1. Filter-Based Feature Selection

Filter-based feature selection is a method that is independent of the model and occurs before model training. Its goal is to select the most representative features through statistical analysis or correlation assessment. In malicious document detection, we can select the most relevant features by calculating the correlation between features and labels.

The advantage of this method is its simplicity and intuitiveness. By quickly selecting features with high relevance to the target classification task, it can reduce data dimensionality to some extent. However, filter-based feature selection has a clear issue: it may overlook the interrelationships between features. In complex high-dimensional data, the redundancy and correlation between features increase, leading to significant issues with the curse of dimensionality and reducing the model's generalization ability. Therefore, in the filter-based feature selection stage, a careful balance between simplicity and the complex relationships between features is needed to improve model performance in high-dimensional space.[6]

3.1.2. Wrapper-Based Feature Selection

Wrapper-based feature selection embeds the feature selection problem into model training, evaluating feature importance based on model performance. In malicious document detection, this means that feature selection is closely related to specific models, such as decision trees, support vector machines, etc.

Compared to filter-based methods, wrapper-based methods can more comprehensively consider the relationships between features, as they directly depend on the model's performance. This helps capture complex interactions between features and improves the model's generalization ability in high-dimensional data. However, the computational cost of wrapper-based feature selection is usually high. Repeated model training to assess the impact of each feature may lead to overfitting the training data.

In malicious document detection, choosing the appropriate model is crucial for the success of wrapper-based feature selection. Therefore, researchers need to balance computational cost and model performance to find the most suitable wrapper-based feature selection method for a specific task.

3.1.3. Embedded Feature Selection

Embedded feature selection embeds the feature selection process into the model's training. This method determines the weights and importance of features through the model's own learning. Common embedded methods in machine learning include L1 regularization, feature importance in decision trees, etc.

In the field of malicious document detection, embedded methods provide a more comprehensive perspective for feature selection, as they fully consider the complex relationships between features. By dynamically adjusting feature weights during training, the model can adaptively learn and emphasize the most predictive features. However, embedded feature selection also requires proper tuning to avoid overfitting the training data.

In high-dimensional feature space, embedded feature selection offers an effective means for malicious document detection tasks. By highlighting the most important features, the model can more accurately capture the key characteristics of malicious documents, improving overall performance.

In summary, the overview of high-dimensional feature selection deeply analyzes the three main feature selection techniques—filtering, wrapping, and embedding—laying the groundwork for the detailed introduction of the proposed high-dimensional feature selection strategy based on machine learning. When facing the challenges of high-dimensional data, choosing the appropriate feature selection technique is crucial.

3.2. Proposed Feature Selection Strategy

3.2.1. Strategy Overview

The high-dimensional feature selection strategy proposed in this paper aims to overcome the challenges of the high-dimensional feature space faced in malicious document detection. This strategy combines the advantages of filtering, wrapping, and embedding feature selection techniques, refining the feature set in three stages: information gain filtering, recursive feature elimination, and embedded feature selection, to obtain the optimal and distinctive feature set.

Initially, the strategy uses the information gain filtering stage to preliminarily select features that contribute significantly to malicious document classification while maintaining key information. Then, through recursive feature elimination, we further refine feature selection, eliminating features with minimal contribution to model performance to achieve a more compact feature set. Finally, embedded feature selection is employed, adjusting the weights of features during the model training process to ensure that the model can adaptively learn and emphasize the most predictive features.

This multi-stage feature selection process allows the strategy to fully consider the complexity of the high-dimensional feature space while effectively reducing computational complexity, providing a feasible solution for the malicious document detection task.

3.2.2. Information Gain Filtering

In the initial filtering stage of the strategy, we use information gain to assess the contribution of features to the classification of malicious documents. Information gain is an effective measure that reflects the degree to which uncertainty is reduced after introducing a feature, helping us find the most distinctive features.

The core idea of information gain filtering is to select features with higher information gain by analyzing the information gain of each feature for the document classification task. The goal of this stage is to preliminarily select a set of features significant for malicious document classification while maintaining key information. Setting an information gain threshold helps us focus on features that best differentiate document categories, laying the foundation for subsequent feature selection processes.

Through information gain filtering, the strategy can quickly and targetedly narrow down the range of the feature set, providing strong support for the next more refined feature selection steps. The effectiveness of this step is directly related to the overall performance improvement of the strategy.

3.2.3. Recursive Feature Elimination

Following information gain filtering, we introduce recursive feature elimination to further refine feature selection. The goal of this stage is to gradually eliminate features with minor contributions to model performance, obtaining a more compact yet still highly discriminative feature set.

Recursive feature elimination iteratively trains the model and removes features with less contribution, ensuring that the final selected feature set maintains high discriminative power while avoiding redundancy. In high-dimensional feature space, this process is crucial for improving model generalization ability. By gradually eliminating unnecessary features, we can obtain a more refined feature set, thereby enhancing the efficiency and performance of the model.

The introduction of recursive feature elimination makes the overall strategy more targeted and

adaptable, further enhancing the performance of the malicious document detection model while maintaining key features. The purpose of this step is to ensure that the final selected features can effectively differentiate documents and help reduce dimensions to address the challenges of high-dimensional feature space.

3.2.4. Embedded Feature Selection

The final stage is embedded feature selection, which occurs during the model's training process. This stage aims to ensure that the model can adaptively learn and emphasize the most predictive features. Specific embedded methods may include L1 regularization, ensuring that the model appropriately focuses on certain high-dimensional features during learning.

Embedded feature selection adjusts the weights of features during the model training process, making the model more focused on the most predictive features for malicious document classification tasks. Compared to the previous two stages, embedded methods more comprehensively consider the complex relationships between features and have stronger adaptability.

The key to this process is to maintain high model performance while reducing dimensionality, thereby overcoming the challenges in the high-dimensional feature space. The introduction of embedded feature selection is an essential part of the strategy, providing the malicious document detection model with a final, refined feature subset to enhance the model's generalization ability.

Through this comprehensive feature selection strategy, we aim to fully utilize the advantages of various techniques, considering the relationships between features and improving model adaptability through embedded methods. This is expected to achieve more accurate and efficient malicious document detection in high-dimensional feature spaces.

3.3. Empirical Study Design

To validate the effectiveness of the proposed high-dimensional feature selection strategy based on machine learning, we designed a series of empirical studies aimed at comprehensively evaluating the strategy's performance and applicability in malicious document detection tasks.

First, we selected multiple public datasets containing different types of malicious document samples, ensuring the generalizability of the experimental results. This choice helps to verify the effectiveness of the proposed strategy in handling diverse threats, making the experimental results more universally applicable and reliable.

Next, we compared the performance of models with and without the high-dimensional feature selection strategy. By evaluating metrics such as accuracy and recall rate, we could deeply understand the applicability of the proposed strategy in different scenarios. This comparison not only focuses on the accuracy of detection but also considers the adaptability to unknown threats, providing a comprehensive evaluation of the proposed strategy's overall performance.

In the experimental design, we paid special attention to factors relevant to practical applications, such as the model's training time and computational resource consumption. This helps to comprehensively evaluate the feasibility of the proposed strategy in large-scale network environments. By considering these factors, we can ensure that the proposed strategy is not only effective theoretically but also practical in real-world applications.

Through a series of empirical studies, we aim to validate that the proposed strategy can maintain efficient detection performance while reducing computational complexity. This will provide reliable support for the practical application of the strategy and lay the groundwork for further optimization and improvement.

4. Experimental Results and Discussion

4.1. Analysis of Experimental Results

In this section, we present the empirical study results of the high-dimensional feature selection strategy based on machine learning in detail and compare the performance of models with and without this strategy. The abundant experimental data highlight the significant advantages of this strategy in improving the accuracy of malicious document detection and reducing computational complexity.

4.1.1. Improving Detection Accuracy

The experimental results clearly show that models using the high-dimensional feature selection strategy achieved higher accuracy in malicious document detection tasks. With a carefully chosen set of features, models were more capable of capturing key characteristics of malicious documents, thereby improving classification accuracy. Comparing the accuracy metrics of the experimental and control groups, we observed a significant performance improvement. This indicates that strategic selection in the high-dimensional feature space allows models to more accurately differentiate malicious documents, thus enhancing overall detection effectiveness.

4.1.2. Reducing Computational Complexity

The experimental results further show that models using the high-dimensional feature selection strategy exhibit superior performance in computational complexity. By reducing the dimensionality of the feature space, the training and prediction time of models significantly decreased, while consuming fewer computational resources. This offers a more efficient solution for practical applications in large-scale network environments. The advantage of reducing computational complexity is reflected not only in time efficiency but also potentially in lowering the hardware costs required for maintaining and deploying models, making it a more attractive option for practical applications.

4.1.3. Validating Applicability

Moreover, the experimental results also validate the applicability of this strategy in different scenarios. Experiments on multiple public datasets ensured the generalizability of the strategy for different types of malicious documents. This empirical study provides strong support for the feasibility of the strategy in practical applications.

In summary, the analysis of experimental results demonstrates the significant advantages of the proposed high-dimensional feature selection strategy based on machine learning in improving detection accuracy and reducing computational complexity. This provides an effective and practical solution for the field of malicious document detection, ensuring more reliable cybersecurity in practical applications.

4.2. Robustness Analysis of the Strategy

We conducted an in-depth analysis of the robustness of the proposed strategy, introducing noise and perturbations to simulate real-world interferences, and evaluated the strategy's performance in these scenarios.

4.2.1. Experiments with Noise Introduction

The experimental results indicate that even in the presence of some degree of noise, the proposed strategy maintained good performance. With the introduction of noise, the model's classification of malicious documents remained highly accurate, demonstrating a certain level of robustness to

disturbances. This proves that in real network environments, where data collection and transmission may introduce uncertainties, the strategy can still effectively counteract the impact of noise and maintain high detection accuracy.

4.2.2. Robustness Analysis under Perturbations

By introducing perturbations, we further examined the strategy's performance in the face of minor changes in model inputs. The results showed that the strategy still maintained relatively stable performance even when slight perturbations occurred in the inputs. This indicates that the proposed feature selection strategy is robust to minor changes in inputs and can handle data fluctuations in practical applications. This is particularly crucial in the field of network security, as the dynamic nature of network environments can lead to changes in input data, while the strategy's robustness ensures efficient detection performance in such situations.

4.2.3. Practical Implications of Robustness

Overall, the experimental results and robustness analysis validate the effectiveness and practicality of the high-dimensional feature selection strategy based on machine learning in malicious document detection tasks. The strategy not only achieved significant results under idealized experimental conditions but also performed excellently in the face of real-world noise and perturbations. This provides strong support for future research and practical applications, emphasizing the importance of feature selection in high-dimensional data environments. The robustness of the strategy means it has stronger adaptability and generalizability, able to perform stably in complex and variable network environments.

5. Conclusion

Through this research, we have validated the positive impact of the high-dimensional feature selection strategy based on machine learning in malicious document detection. This strategy not only improves model performance but also reduces the computational cost, offering an effective solution for the field of malicious document detection. Future research could further explore combinations of different feature selection algorithms and the application of this strategy in large-scale network environments.

Acknowledgement

"Research on Malicious Document Detection System Based on High-Dimensional Features in Machine Learning" supported by the Education Department of Hainan Province, project number Hnky2021-62.

References

[1] Huang Kun. Visualization Detection of Malicious Documents Based on Deep Learning [J]. Electronic Measurement Technology. 2022, 45(18): 126-133.
[2] Liao Jinzhi. Cross-document False Information Detection Based on Comparative Graph Learning [J]. Computer Science. 2023 (12): 9.
[3] Zhang Rong. Machine Learning-based Detection of ACM Performance Degradation Failures [J]. Aviation Maintenance and Engineering. 2023 (11): 40-42.
[4] Yue Xin. Application of Hyperspectral Imaging and Short Video Imaging Combined with Machine Learning in Detecting the Consistency of Fireproof Coatings [J]. Coatings Industry. 2023 (12): 8.
[5] Qin Chuandong. Multi-Strategy Hybrid Artificial Bee Colony Algorithm for High-Dimensional Feature Selection of Microarrays [J]. Journal of System Simulation. 2023, 35(03): 515-524.
[6] Ma Yunpeng. Machine Learning-based Estimation of Microbial Dissolved Organic Carbon Content [J]. Advances in Biotechnology. 2023, 13(04): 645-653.