Algorithms for Spam Detection

Subjects: Computer Science, Artificial Intelligence

Contributor: Mohammad Tubishat , Feras Al-Obeidat , Ali Safaa Sadiq , Seyedali Mirjalili

Spam emails have become a pervasive issue, as internet users receive increasing amounts of unwanted or fake emails. To combat this issue, automatic spam detection methods have been proposed, which aim to classify emails into spam and non-spam categories. Machine learning techniques have been utilized for this task with considerable success.

cybersecurity optimization feature selection algorithms

1. Introduction

With the increasing use of the internet and online social networks (OSNs) applications, communication and the exchange of information among users have similarly increased. Along with this increased communication comes the problem of spam, which is an issue that users of these applications frequently face. One of the most common forms of spam is unsolicited emails or spam emails, which fill up email inboxes and take time for users to check and delete [1][2]. The problem of spam is not limited to just email but also affects other network applications [3]. For instance, users of social networking sites often receive unwanted messages or comments from fake accounts or spammers. Such messages can be annoying, and harmful, and can lead to privacy breaches, identity theft, and financial losses [4].

To address this problem, spam filtering software is developed and employed to detect and remove spam emails [5]. However, these filtering systems may not be accurate all the time and may mistakenly classify legitimate emails as spam [6]. This can lead to users missing out on important information or communication, such as important emails for job offers, contracts to sign, important appointments, etc. Additionally, spammers can use various tactics to bypass these filters and send malicious emails that are designed to deceive users into disclosing their personal information, passwords, or financial details. Therefore, a robust and accurate spam detection method is necessary to detect and prevent these threats [7].

2. Algorithms for Spam Detection

Over the last few years, optimization algorithms have been widely used for feature selection in spam email detection methods. Numerous search studies have proposed several optimization-based spam detection approaches, and their efficiency and powers have been widely analysed. For example, in Sokhangoee and [8], a spam detection method based on association-rule mining and a genetic algorithm is proposed. The method achieved high accuracy in detecting spam emails, though it was suffering from high computational complexity. In

contrast, [9] proposed a spam detection method based on the combination of the Harris Hawks Optimizer (HHO) and the KNN classifier. The method has demonstrated promising results in terms of accuracy and processing time. Though, the HHO algorithm by its nature is heavily dependent on the random initialization of its parameters, which may impact its stability and reproducibility.

On the other hand [10] introduced a spam detection method based on the Horse Herd Optimization Algorithm (HOA) with a KNN classifier. Their method gained high accuracy in detecting spam emails; nonetheless, its performance could be heavily impacted by the sensitivity of the HOA algorithm, which comes from the nature of its parameter settings.

Another attempt [11] proposed the use of the Symbiotic Organisms Search (SOS) algorithm in the spam email detection mechanism. Their proposed approach has demonstrated high accuracy in detecting spam emails and has performed well in contrast with other optimization-based approaches. Yet, the introduced computational cost could be relatively high, which bounds its feasibility with a large-scale spam detection problem.

On the other hand, the authors in [12] suggested the use of the sine–cosine algorithm (SCA) in detecting spam emails. The proposed approach has performed well in terms of accuracy and processing time. However, the performance could be limited by the nature of the SCA algorithm's sensitivity, due to the nature of its parameter settings. Hence, such a method will not be a reliable option, especially when it comes to the highly sensitive nature of the detection process of spam email filtering mechanisms.

The authors in [13] introduced the Water Cycle Optimization (WCO) algorithm in conjunction with Simulated Annealing (SA) to be used in detecting spam emails. Though their proposed method has demonstrated high accuracy in detecting spam emails, its computational complexity was relatively high.

From the presented methods and approaches, it can find out the potential use of optimization algorithms in spam email detection. Though, their performance varies depending on specific algorithmic features, parameter settings, and computational complexity. Additional research is therefore needed as a way to develop more efficient and effective optimization-based spam detection methods.

In **Table 1** a comparison of some of the common nature-inspired metaheuristic algorithms based on their population, individual, and optimization strategies is listed. The evolutionary types of algorithms, such as genetic algorithms and differential evaluation strategies, generally depend on the concept of natural selection to optimize solutions over a population of individuals, while swarm-based algorithms, such as particle swarm optimization and firefly algorithms, mimic the collective behaviour of social swarms to optimize the given solutions. Physical-based algorithms, such as simulated annealing and harmony search, are inspired by physical phenomena like thermal energy and musical harmony to optimize solutions. Other metaheuristic algorithms, such as grey wolf optimization [14], the artificial bee colony algorithm, and the imperialist competitive algorithm, draw inspiration from various sources to optimize solutions.

Table 1. Comparison of the nature-inspired metaheuristic algorithms.

Nature-Inspired Metaheuristics				
	Evolutionary Algorithms	Swarm-Based Algorithms	Physical-Based Algorithms	Other Metaheuristics
Population	Genetic Algorithms	Particle Swarm Algorithms	Simulated Annealing	Grey Wolf Optimization
Individual	Differential Evaluation Strategies	Firefly Algorithms	Harmony Search	Artificial Bee Colony Algorithm
Optimization Strategy	Evolutionary Programming	Ant Colony Optimization Algorithm	Memetic Algorithms	Imperialist Competitive Algorithm

It is important to note that the choice of a metaheuristic algorithm is heavily dependent on the specific optimization problem at hand. For instance, swarm-based algorithms are often used for optimization problems that require the exploration of a large search space, while physical-based algorithms are often used for optimization problems that require the optimization of a continuous function. In addition, hybrid metaheuristic algorithms that combine different techniques from different categories have been proposed to achieve better performance in optimization problems.

In order to demonstrate some of the key analysis aspects that could be used in comparing optimization techniques, below are some analysis points that could be used to highlight the competency of the related works:

Performance comparison: In addition to listing the strengths and weaknesses of each algorithm, this comparison can be performed based on various metrics such as accuracy, precision, recall, F1 score, etc. The comparison can also be performed on different datasets to evaluate the generalizability of the algorithms.

Impact of feature selection: Many of the algorithms mentioned in the related works section use feature selection techniques to improve the accuracy of spam detection. This analysis could demonstrate the impact of feature selection on the performance of the algorithms. This analysis could also include a comparison of the performance of algorithms with and without feature selection and compare different feature selection techniques.

- Analysis of false positives and false negatives: False positives and false negatives are common errors in spam detection. An analysis of the false positives and false negatives generated by each algorithm could be used on each of these algorithms to compare and contrast them. This analysis could help identify the specific types of emails that are misclassified by each algorithm and suggest improvements to reduce these errors.

- Robustness analysis: The robustness of the algorithms could be analysed by testing their performance under different scenarios such as varying spam densities, different types of spam, and changes in the email dataset. This analysis could help evaluate the generalizability of the algorithms and identify scenarios where they may not perform well.

- Comparison with traditional spam detection methods: Such a comparison could compare the performance of the optimization algorithms with traditional rule-based and content-based spam detection methods. This comparison could help evaluate the effectiveness of optimization algorithms in improving the accuracy of spam detection.

- Analysis of computational efficiency: Optimization algorithms can be computationally expensive, especially when dealing with large datasets. The computational efficiency of each algorithm could be analysed and compared with their run times on different datasets. This analysis could help identify the most efficient algorithms and suggest improvements to reduce their computational cost.

- On the other hand, the DO is a relatively new optimization algorithm that has been applied to various optimization problems, including feature selection and classification tasks, which has the potential to be used for spam detection. As with any other optimization algorithm, DO has some limitations, which are listed as follows:

- Premature Convergence: DO tends to converge prematurely to local optima, which can result in suboptimal solutions [15]. This is a common problem in many optimization algorithms and the DO algorithm is no exception.

- Sensitivity to Initialization: DO's performance can be sensitive to the initial population's quality and diversity [16]. Poor initialization can lead to premature convergence, while good initialization can improve the algorithm's performance.

- Lack of Diversity: DO does not have mechanisms to maintain population diversity, which can cause premature convergence and limit the algorithm's exploration capabilities [17].

- Limited Search Space Exploration: DO's search capabilities are limited, as it only explores a small portion of the search space at each iteration. This can result in suboptimal solutions and can make it difficult to find the global optimum [18].

- Computational Complexity: DO's computational complexity can be high, particularly for large-scale problems. The algorithm involves evaluating fitness functions, which can be computationally expensive, and the algorithm's complexity can increase with the problem's dimensionality [19].

- Lack of Theoretical Analysis: DO's theoretical analysis is still limited, and there are few theoretical guarantees of its convergence and performance under different conditions. This makes it difficult to understand the algorithm's behaviour and to design effective parameter settings [20].

In summarizing the performance evaluation of the DO algorithm, it has exhibited encouraging outcomes in certain applications; however, researchers need to acknowledge its limitations and drawbacks when considering its application to their specific optimization problems. To enhance the algorithm's effectiveness, researchers should investigate strategies to address and overcome these limitations.

While many optimization techniques have been utilized in the literature for feature selection in spam email detection, the No Free Lunch Theorem (NFL) [21] suggests that no single solution can be applied to all problems and outperform all other algorithms. Hence, researchers continue to investigate the use of the most recent optimization algorithms for spam email detection, including DO.

Table 2 provides a summary of several optimization algorithms, including the Particle Swarm Optimization (PSO), the Genetic Algorithm (GA), Ant Colony Optimization (ACO), the Artificial Bee Colony (ABC), Hill Climbing, Simulated Annealing, and Tabu Search. The strengths and weaknesses of each algorithm are listed, as well as their effectiveness in email spam detection. The table suggests that PSO, GA, ACO, and ABC have shown promising results in email spam detection, particularly for feature selection and email classification. However, each algorithm has its limitations and requires careful parameter tuning for optimal performance. Hill Climbing, Simulated Annealing, and Tabu Search have been used successfully for email classification but may not be as effective as other optimization algorithms for feature selection. Overall, the table provides a useful reference for researchers to choose an appropriate optimization algorithm for their email spam detection problem based on their specific requirements and constraints.

Table 2. Summary of optimization algorithms application in email spam detection.

Optimization Algorithm	Description	Strengths	Weaknesses	Effectiveness in Email Spam Detection
Particle Swarm Optimization (PSO)	A population-based optimization algorithm that involves particles moving around in the search space to find the best solution.	Good for feature selection, can handle high-dimensional data, easy to implement.	Can become stuck in local optima, sensitive to parameter settings.	Has shown promising results in email spam detection, particularly for feature selection and email classification.
Genetic Algorithm (GA)	A population-based optimization algorithm that involves creating a population of potential solutions and then applying selection, crossover, and mutation operations to evolve the population over generations.	Can handle non-linear and non-convex problems and can find multiple optimal solutions.	Can be slow, requires careful parameter tuning, and may suffer from premature convergence.	Has been used successfully for email spam detection, particularly for email classification.
Ant Colony Optimization (ACO)	An optimization algorithm that uses pheromone trails to guide the search process.	Good for feature selection, can handle high-dimensional data, and can	Can be slow, sensitive to parameter settings, and may suffer from	Has shown promising results in email spam detection, particularly for feature selection

Optimization Algorithm	Description	Strengths	Weaknesses	Effectiveness in Email Spam Detection
		find global optima.	premature convergence.	and email classification.
Artificial Bee Colony (ABC)	An optimization algorithm that involves employed bees, onlooker bees, and scout bees to explore the search space.	Good for finding global optima, easy to implement.	Can be slow, sensitive to parameter settings, and could suffer from premature convergence.	Has been used successfully for email spam detection, particularly for email classification.
Hill Climbing	A local search algorithm that iteratively improves the current solution by making small changes to it.	Simple and fast, can handle large datasets.	Can become stuck in local optima and could not find global optima.	Has been used successfully for email classification but may not be as effective as other optimization algorithms for feature selection.
Simulated Annealing	An optimization algorithm that starts with a high “temperature” and then gradually decreases it to find the best solution.	Able to find global optima, and manage noisy data.	Can be slow, and sensitive to parameter settings.	Has been used successfully for email classification but may not be as effective as other optimization algorithms for feature selection.
Tabu Search	A metaheuristic algorithm that is based on the concept of intensification and diversification.	Able to solve non-linear and non-convex problems, also finding global optima.	Can be slow and requires careful parameter tuning.	Has been used successfully for email classification but may not be as effective as other optimization algorithms for feature selection.

A

3).

3. Doshi, J.; Parmar, K.; Sanghavi, R.; Shekokar, N. A comprehensive dual-layer architecture for phishing and spam email detection. Comput. Secur. 2023, 133, 103378.
4. Back, S.; LaPrade, J. Cyber-Situational Crime Prevention and the Breadth of Cybercrimes among Higher Education Institutions. Int. J. Cybersecur. Intell. Cybercrime 2020, 3, 25–47.
5. Saidani, N.; Adi, K.; Allili, M.S. A semantic-based classification approach for an enhanced spam detection. Comput. Secur. 2020, 94, 43–56.
6. Khandelwal, Y.; Bhargava, R. Spam filtering using AI. In Artificial Intelligence and Data Mining Approaches in Security Frameworks; Wiley: Hoboken, NJ, USA, 2021; pp. 87–99.
7. Amin, M.; Al-Obeidat, F.; Tubaishat, A.; Shah, B.; Anwar, S.; Tanveer, T.A. Cyber security and beyond: Detecting malware and concept drift in AI-based sensor data streams using statistical

techniques. Comput. Electr. Eng. 2023, 108, 108702.

8. Sokhangoee, Z.F.; Rezapour, A. A novel approach for spam detection based on association rule mining and genetic algorithm. Comput. Electr. Eng. 2022, 97, 107655.
9. Mashaleh, A.S.; Ibrahim, N.F.B.; Al-Betar, M.A.; Mustafa, H.M.J.; Yaseen, Q.M. Detecting Spam Email with Machine Learning Optimized with Harris Hawks optimizer (HHO) Algorithm. Procedia Comput. Sci. 2022, 201, 659–664.
10. Hosseinalipour, A.; Ghanbarzadeh, R. A novel approach for spam detection using horse herd optimization algorithm. Neural Comput. Appl. 2022, 34, 13091–13105.
11. Mohammadzadeh, H.; Gharehchopogh, F.S. Feature Selection with Binary Symbiotic Organisms Search Algorithm for Email Spam Detection. Int. J. Inf. Technol. Decis. Mak. 2021, 20, 469–515.
12. Pashiri, R.T.; Rostami, Y.; Mahrami, M. Spam detection through feature selection using artificial neural network and sine–cosine algorithm. Math. Sci. 2020, 14, 193–199.
13. Al-Rawashdeh, G.; Mamat, R.; Rahim, N.H.B.A. Hybrid Water Cycle Optimization Algorithm with Simulated Annealing for Spam E-mail Detection. IEEE Access 2019, 7, 143721–143734.
14. Mirjalili, S.; Mirjalili, S.M.; Lewis, A. Grey Wolf Optimizer. Adv. Eng. Softw. 2014, 69, 46–61.
15. Bhatnagar, V.; Sharma, V. Comparative Study of Dandelion and Firefly Algorithms for Parameter Estimation of a Dynamic System. ISA Trans. 2020, 102, 121–131.
16. Wang, Z.; Li, S.; Wang, Y. Improved dandelion algorithm for global optimization problems. IEEE Access 2020, 8, 30799–30810.
17. Zhu, H.; Liu, G.; Zhou, M.; Xie, Y.; Kang, Q. Dandelion algorithm with probability-based mutation. IEEE Access 2019, 7, 97974–97985.
18. Javed, M.A.; Al-Rifaie, M.M. A comparative study of the Dandelion algorithm with recent swarm intelligence algorithms. Appl. Soft Comput. 2019, 84, 105712.
19. Namin, A.S.; Hosseinabadi, S.; Namin, A.S. A novel hybrid Dandelion algorithm with biogeography-based optimization for solving the economic emission load dispatch problem. J. Clean. Prod. 2020, 273, 122824.
20. Xu, G.; Wang, Z.; Sun, L. Hybrid dandelion algorithm for global optimization problems. Soft Comput. 2020, 24, 10903–10915.
21. Wolpert, D.H.; Macready, W.G. No free lunch theorems for optimization. IEEE Trans. Evol. Comput. 1997, 1, 67–82.