ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940
www.ijsem.org
Vol 12, Issuse . 3 Aug 2021

A BI-OBJECTIVE HYPER-HEURISTIC SUPPORT VECTOR MACHINES FOR BIG DATA CYBER-SECURITY

Dr.A.Ramaswami Reddy,

ABSTRACT:

Cyber security in the context of big data is known to be a critical problem and presents a great challenge to the research community. Machine learning algorithms have been suggested as candidates for handling big data security problems. Among these algorithms, support vector machines (SVMs) have achieved remarkable success on various classification problems. However, to establish an effective SVM, the user needs to dene the proper SVM configuration in advance, which is a challenging task that requires expert knowledge and a large amount of manual effort for trial and error. In this paper, we formulate the SVM configuration process as a bi-objective optimization problem in which accuracy and model complexity are considered as two conflicting objectives. We propose a novel hyper-heuristic framework for bi-objective optimization that is independent of the problem domain. This is the first time that a hyper-heuristic has been developed for this problem. The proposed hyper-heuristic framework consists of a high-level strategy and low-level heuristics. The high-level strategy uses the search performance to control the selection of which low-level heuristic should be used to generate a new SVM configuration. The low-level heuristics each use different rules to effectively explore the SVM configuration search space. To address bi-objective optimization, the proposed framework adaptively integrates the strengths of decomposition- and Pareto-based approaches to approximate the Pareto set of SVM configurations. The effectiveness of the proposed framework has been evaluated on two cyber security problems: Microsoft malware big data classification and anomaly intrusion detection. The obtained results demonstrate that the proposed framework is very effective, if not superior, compared with its counterparts and other algorithms.

Keywords:SVM, pareto, object, cyber security problems.

1. INTRODUCTION:

Modern digital information era has created the space for high volume of data to be generated and stored by the advanced technologies and Internet of Things (IoT) [1]. This rapid growth of the Internet data has also exponentially increased the frequency of cyber-attacks. The cyber-attacks cause extensive damages to the networks and hence to tackle them the cyber security systems have been designed and installed. Cyber security techniques and processes are assigned with the role of thwarting the illegal cyber-attacks to protect the computers and networks from the cyber damages [2]. They perform the major function of protecting the shared information for improving decision making; detecting the vulnerable attacks in applications; prevent unauthorized accessing of networks and secure the confidential network information [3]. Most of the larger companies have their own cyber security network while other organizations make use of such solutions from security organizations like

Professor, Dept. of CSE,
Malla Reddy Engineering College (Autonomous), Secunderabad, Telangana State

ISSN2454-9940
www.ijsem.org
Vol 12, Issuse . 3 Aug 2021

Accenture, IBM, CISCO, etc. [4]. Recent cyber security solutions have inclined more towards the monitoring of network and Internet traffic to identify and avert the bad actions [5]. This is entirely different from the traditional cyber security solutions which focus only on the detection of bad signatures for unauthorized access. While the traditional systems were aimed at detecting the malware by scanning the incoming traffic against the malware signatures, they are relatively weaker with detecting only limited threats [6]. These traditional techniques including the intrusion detection, firewalls and anti-virus software have become ineffective in tackling the hackers as the attack strategies are highly destructive than the older versions [7]. In addition to this, the presence of big data has increased the critical condition as gigabytes of data are transferred between each node of the computer networks; making the hackers job of entering the networks very easier and cause severe damage without getting traced [8]. The big data problems are majorly due to the organizations providing access to their data networks allowing the partners and consumers to access all data and making it vulnerable to the cyber-attacks. Similarly, the big data has also increased the skills of hackers to evade the traditional security systems. Also, the big data has made it difficult to identify the attacks when initiated and the attack is only known after the damage is done to the hardware and software components [9]. To address these security threats linked to the big data, the big data analytics can be used for cyber security analytics by employing the big data techniques to evade the cyber-attacks [10]. Based on this concept, many organizations have started to remodel the cyber security systems [11]. As said before, the machine learning algorithms have been utilized extensively for this process with the SVM emerging as the front-runner.

2. LITERATURE SURVEY

Many researchers have focused on developing efficient cyber security solutions using big data analytics. Some of the most recent techniques are discussed in this section. Dovom et al. [13] presented a fuzzy pattern tree method for malware detection of big data IoT. This method transmutes the Op-codes into vector space and applies y fuzzy and fast fuzzy pattern tree to detect the malwares. The results provided high degree of accuracy in categorization with about 97% for Kaggle dataset and more than 93.13% for Ransomware dataset. Shamshirband and Chronopoulos [14] developed high performance-ELM based malware detection method which provided accuracy of 95.92%. However, this model does consider only three features for malware detection and hence needs to be improved. Zhong and Gu [15] proposed a multi-level deep learning system detecting the malwares. This system organizes multiple deep learning models using the tree structure and each tree focuses on specific data distribution of particular malware group. Experimental results showed high accuracy of malware detection using this system but the major drawback is the high computation time. Ju et al. [16] proposed a big data analytics framework for targeted cyber-attacks detection from the heterogeneous noisy data. This approach utilized different heterogeneous data and correlated them to identify the malicious nodes. It provides highly accurate cyber-attack detection but it considers only limited features and does not include the human perception in attack detection. Venkatraman et al. [17] introduced a hybrid deep learning image-based analysis model for detecting the cyber-attacks. This hybrid model helps in detecting suspicious behavior of systems and also visualizes the malware classification. This model achieves high accuracy of malware detection with less computational cost. Calvert and Khoshgoftaar [18] used the big data sampling to produce varying class distributions for the detection of slow HTTP DoS attacks. This approach is based on the legitimate traffic monitoring of the system to detect the attacks using Random forest as optimal learning

ISSN2454-9940
www.ijsem.org
Vol 12, Issuse . 3 Aug 2021

algorithm. This approach provided results of AUC value 0.99904 for the attack detection. However, only the AUC metric is estimated and this causes statistical inconsequential decisions. Mao et al. [19] presented a spatio-temporal approach to detect the malwares based on the big data characteristics of the cloud systems. This approach devised a graph based semi-supervised learning algorithm for detecting the attacks based on the spatial and temporal features of the data distributions. Experimental results provided better detection rate of malwares in less computation time. But in this approach there is an upper bound on the recall to malware detection based on the file co-occurrence in end hosts. Martín et al. [20] introduced MOCDroid using multi-objective evolutionary classifier detecting the malwares in Android. This approach utilizes SPEA2, a multi-objective genetic algorithm, to select groups of import terms to determine the malware nodes. Empirical results proved that this approach has high accuracy and reduced number of false positives; but the approach considers only few objectives. Gupta and Rani [21] proposed machine learning based big data framework for zero-day malware detection. The identification of attacks is performed using classification algorithms in which the random forests provided higher accuracy. However, the larger dataset used for evaluation makes the detection process very slow. Wassermann and Casas [22] developed BIGMOMAL method using big data analytics and supervised-machine-learning for mobile malware detection. This approach detected the malware in running apps with high accuracy but the approach suffers from concept drift problem. From the literature, it can be understood that the machine learning algorithms can provide better classification in the detection of malwares. However, it is also inferred that certain classifiers are only suitable for particular type of datasets. This leads to the necessity of the designing better configuration of the machine learning algorithms which could provide highly accurate malware detection with less computation time and higher efficiency.

3. METHODOLOGY

EXISITNG SYSTEM:

SVMs are a class of supervised learning models that have been widely used for classification and regression SVMs are based on statistical learning theory and are better able to avoid local optima than other classification algorithms. An SVM is a kernel-based learning algorithm that seeks the optimal hyper plane. The kernel learning process maps the input patterns into a higher-dimensional feature space in which linear separation is feasible. The existing kernel functions can be classified as either local or global kernel functions. Local kernel functions have a good learning ability but do not have good generalization ability. By contrast, global kernel functions havegood generalization ability but a poor learning ability. For example, the radial kernel function is known to be a local function, whereas the polynomial kernel function is a global kernel function. The main challenge lies in determining which kernel function should be used for the current problem instance or the current decision point. This is because the kernel selection process strongly depends on the distribution of the input vectors and the relationship between the input vector and the output vector (predicted variables). However, the feature space distribution is not known in advance and may change during the course of the solution process, especially in big data cyber security. Consequently, different kernel functions may work well for different instances or in different stages of the solution process and kernel selection may thus have a crucial impact on SVM performance. To address this issue, in this work, we use multiple kernel functions to improve the accuracy of our algorithm and avoid the shortcomings of using a single kernel function.

PROPOSED SYSTEM:

The proposed hyper-heuristic framework for configuration selection is shown in Figure 2. It has two levels: the high- level strategy and the low-level heuristics. The high-level strategy operates on the heuristic space instead of the solution space. In each iteration, the high-level

ISSN2454-9940
www.ijsem.org
Vol 12, Issuse . 3 Aug 2021

strategy selects a heuristic from the existing pool of low-level heuristics, applies it to the current solution to produce a new solution and then decides whether to accept the new solution. The low level heuristics constitute a set of problem-specific heuristics that operate directly on the solution space of a given problem. To address the bi-objective optimization problem, we propose a population-based hyper-heuristic framework that operates on a population of solutions and uses an archive to save the non-dominated solutions. The proposed framework combines the strengths of decomposition- and Pareto (dominance) - based approaches to effectively approximate the Pareto set of SVM configurations. Our idea is to combine the diversity ability of the decomposition approach with the convergence power of the dominance approach. The decomposition approach operates on the population of solutions, whereas the dominance approach uses the archive. The hyper heuristic framework generates a new population of solutions using the old population, the archive, or both the old population and the archive. This allows the search to achieve a proper balance between convergence and diversity. It should be noted that seeking good convergence involves minimizing the distances between the solutions and PF, whereas seeking high diversity involves maximizing the distribution of the solutions along PF. The main components of the proposed hyper-heuristic framework are discussed in the following subsections.

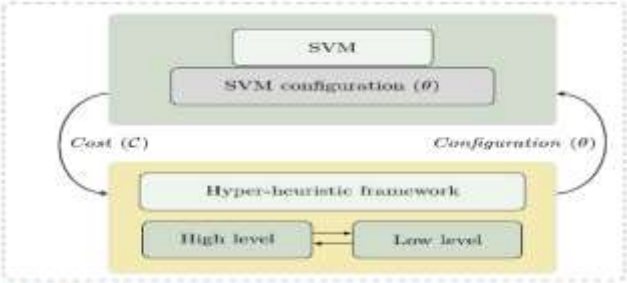


FIGURE 1. The proposed methodology.

In this paper author is describing concept to automatically select SVM optimize parameters by applying multi objective hyper heuristic technique. In this technique two hyper variables such as Low Level Heuristic and High Level strategic can be used to select optimized parameters for SVM. The high-level strategy operates on the heuristic space instead of the solution space. In each iteration, the high level strategy selects a heuristic from the existing pool oflow-level heuristics, applies it to the current solution to produce a new solution and then decides whether to accept the new solution.

HLH will select SVM parameters and then generate a population and then LLH will apply those parameters and then run SVM algorithms and if algorithm give highest accuracy then LLH will accept that parameters as SOLUTION and if not accuracy is optimize then HLH and LLH will keep on generating new population and then evaluate fitness as accuracy. This step continues till highest fitness achieved and that highest fitness parameters will be selected as solution. For each iteration HHSVM will calculate crossover (generate new solution), mutation (select new input values) and fitness (evaluate accuracy). The high fitness value will be selected as best solution. In below screen with comments we are showing HHSVM implementation

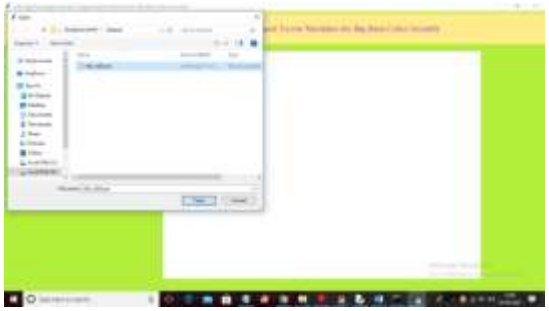
In above screen in selected text we define parameters for HHSVM and in below screen we are using code to perform parameter selection

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940
www.ijsem.org
Vol 12, Issuse . 3 Aug 2021

In above screen read red colour comments to select input for HHSVM and in below screen you can see HHSVM implementation

In above screen click on 'Upload NSL-KDD Dataset' button to upload dataset and to get below screen



In above screen selecting and uploading 'NSL-KDD.txt' dataset and then click on 'Open' button to load dataset and to get below screen

In above screen dataset loaded and I am displaying few records from dataset and in above dataset we can see some values are non-numeric and machine learning will not accept non-numeric values so we need to preprocess those values to assign integer id to each unique non-numeric value

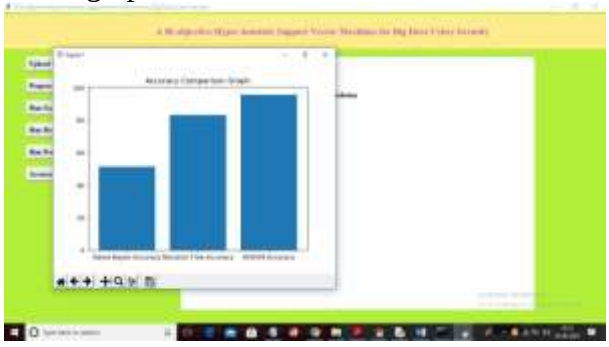
In above screen dataset is preprocessed and dataset contains huge 10137 records and application split dataset into train and test where application using 8109 records for training and 2028 records testing trained ML model accuracy. After train model test records will apply on trained model to perform prediction and then correct prediction percentage will be consider as accuracy. Now train and test dataset is ready and now click on 'Run Gaussian Naïve Bayes Algorithm' button to train Naïve Bayes with above dataset

In above screen Naïve Bayes got 51% accuracy and now click on 'Run Decision Tree Algorithm' button to train decision tree on above dataset

In above screen with decision tree we got 83% accuracy and now click on 'Run Propose HHSVM Algorithm' button to train HHSVM algorithm with optimize parameters and then calculate best fitness accuracy

ISSN2454-9940
www.ijsem.org
Vol 12, Issuse . 3 Aug 2021

In above screen with optimize parameters we got 95% accuracy for HHSVM algorithm and in above accuracy line I am printing all those optimize parameters which helps in getting 95% accuracy and in above screen we can see by applying 'Bi-objective Hyper-heuristic' technique for SVM we got high accuracy and now click on 'Accuracy Comparison Graph' button to get below graph



In above graph x-axis represents algorithm name and y-axis represents accuracy of those algorithms and from above graph we can conclude that HHSVM got high accuracy.

CONCLUSION

In this work, we proposed a hyper-heuristic SVM optimization framework for big data cyber security problems. We formulated the SVM configuration process as a bi-objective optimization problem in which accuracy and model complexity are treated as two conflicting objectives. This bi-objective optimization problem can be solved using the proposed hyper-heuristic framework. The framework integrates the strengths of decomposition- and Pareto-based approaches to approximate the Pareto set of configurations.

REFERANCES

1. Abomhara, M. (2015). "Cyber security and the internet of things: vulnerabilities, threats, intruders and attacks." Journal of Cyber Security and Mobility, 4(1), 65-88.

2. Von Solms, R., & Van Niekerk, J. (2013). "From information security to cyber security." computers & security, 38, 97-102.

3. Probst, C. W., Hunker, J., Bishop, M., & Gollmann, D. (Eds.). (2010). "Insider threats in cyber security" (Vol. 49). Springer Science & Business Media.

4. Moore, T. (2010). "The economics of cybersecurity: Principles and policy options." International Journal of Critical Infrastructure Protection, 3(3-4), 103-117.

5. Choo, K. K. R. (2011). "The cyber threat landscape: Challenges and future research directions." Computers & Security, 30(8), 719-731.

6. Cardenas, A., Amin, S., Sinopoli, B., Giani, A., Perrig, A., & Sastry, S. (2009). "Challenges for securing cyber physical systems." In Workshop on future directions in cyber-physical systems security (Vol. 5, No. 1).

7. Greitzer, F. L., & Frincke, D. A. (2010). "Combining traditional cyber security audit data with psychosocial data: towards predictive modeling for insider threat mitigation." In Insider threats in cyber security (pp. 85-113). Springer, Boston, MA.

8. Hu, J., & Vasilakos, A. V. (2016). "Energy big data analytics and security: challenges and opportunities." IEEE Transactions on Smart Grid, 7(5), 2423-2436.

9. Babiceanu, R. F., & Seker, R. (2016). "Big Data and virtualization for manufacturing cyber-physical systems: A survey of the current status and future outlook." Computers in Industry, 81, 128-137.

10. Mahmood, T., & Afzal, U. (2013). "Security analytics: Big data analytics for cybersecurity: A review of trends, techniques and tools." In 2013 2nd national conference on Information assurance (ncia) (pp. 129-134). IEEE.

11. Kache, F., & Seuring, S. (2017). "Challenges and opportunities of digital information at the intersection of Big Data Analytics and supply chain management." International Journal of

ISSN2454-9940
www.ijsem.org
Vol 12, Issuse . 3 Aug 2021

Operations & Production Management, 37(1), 10-36.

12. Sabar, N. R., Yi, X., & Song, A. (2018). "A bi-objective hyper-heuristic support vector machines for big data cyber-security." IEEE Access, 6, 10421-10431.

13. Dovom, E. M., Azmoodeh, A., Dehghantanha, A., Newton, D. E., Parizi, R. M., & Karimipour, H. (2019). "Fuzzy pattern tree for edge malware detection and categorization in IoT." Journal of Systems Architecture, 97, 1-7.