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

Mr. [1]ASEENABABU SHAIK, Associate Professor

[2]Shreeja, [3]Shama, [4]Shivani, [5]Manideep, [6]Nikhilesh

[2,3,4,5,6]*student, Department Of Computer Science, Siddhartha Institute Of Technology AndSciences*

Narapally –Hyderabad , Telangana

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 variousclassification problems. However, to establish an effective SVM, the user needs to define theproper 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 theproblem 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.

INTRODUCTION

The rapid advancements in technologies and networking such as mobile, social and Internet of Things create massive amounts of digital information. In this context, the termbig data has been emerged to describe this massive amounts of digital information. Big data refers to large and complex datasets containing both structured and unstructured data generated on a daily basis, and need to be analysed in short periods of time. The term big data is different from the big database, where big data indicates the data is too big, too fast,or too hard for existing tools to handle. Big data is commonly described by three characteristics: volume, variety and velocity (aka 3Vs). The 3Vs define properties or dimensions of data where volume refers to an extreme size of data, variety indicates the data was generated from drivers sources and velocity refers to the speed of data creation, streaming and aggregation.

However, the performance of a meta-heuristic method strongly depends on the selected parameters and operators, the selection of which is known to be a very difficult and time-consuming process. In addition, only one kernel is used in most works, and the search is performed over the parameter space of that kernel. This work presents a novel bi- objective hyper-heuristic framework for

SVM configuration optimisation. Hyper- heuristics are more effective than other methods because they are independent of the particular task at hand and can often obtain highly competitive configurations. Our proposed hyper-heuristic framework integrates several key components that differentiate it from existing works to find an effective SVM configuration for big data cyber security. First, the framework considers a bi-objective formulation of the SVM configuration problem, in which the accuracy and model complexity are treated as two conflicting objectives. Second, the framework controls the selection of both the kernel type and kernel parameters as well as the soft margin parameter. Third, the hyper-heuristic framework combines the strengths of decomposition- and Pareto-based approaches in an adaptive manner to find an approximate Pareto set of SVM configurations.

LITERATURE SURVEY

1. *Novel feature extraction, selection and fusion for effective malware familyclassification*

Modern malware is designed with mutation characteristics, namely polymorphism and metamorphism, which causes an enormous growth in the number of variants of malware samples. Categorization of malware samples on the basis of their behaviors is essential for the computer security community, because they receive huge number of malware everyday, and the signature extraction process is usually based on malicious parts characterizing malware families. Microsoft released a malware classification challenge in 2015 with a huge dataset of near 0.5 terabytes of data, containing more than 20K malware samples. The analysis of this dataset inspired the development of a novel paradigm that is effective in categorizing malware variants into their actual family groups. This paradigm is presented and discussed in the present paper, where emphasis has been given to the phases related to the extraction, and selection of a set of novel features for the effective representation of malware samples. Features can be grouped according to different characteristics of malware behavior, and their fusion is performed according to a per-class weighting paradigm. The proposed method achieved a very high accuracy (≈ 0.998) on the Microsoft Malware Challenge dataset.

2. *Efficient string match ing: an aid to bibliographic search. Communications of theACM*

This paper describes a simple, efficient algorithm to locate all occurrences of any of a finite number of keywords in a string of text. The algorithm consists of constructing a finite state pattern matching machine from the keywords and then using the pattern matching machine to process the text string in a single pass. Construction of the pattern matching machine takes time proportional to the sum of the lengths of the keywords. The number of state transitions made by the pattern matching machine in processing the text string is independent of the number of keywords. The algorithm has been used to improve the speed of a library bibliographic search program by a factor of 5 to 10.

3. *A meta-learning approach to automatic kernel selection for support vectormachines*

Appropriate choice of a kernel is the most important ingredient of the kernel-based learning methods such as support vector machine (SVM). Automatic kernel selection is a key issue given the number of kernels available, and the current trial-and-error nature of selecting the best kernel for a given problem. This paper introduces a new method for automatic kernel selection, with empirical results based on classification. The empirical study has been conducted among five kernels with 112 different classification problems, using the popular kernel based statistical learning algorithm SVM. We evaluate the kernels' performance in terms of accuracy measures. We then focus on answering the question:

which kernel is best suited to which type of classification problem? Our meta-learning methodology involves measuring the problem characteristics using classical, distance and distribution-based statistical information. We then combine these measures with the empirical results to present a rule-based method to select the most appropriate kernel for a classification problem. The rules are generated by the decision tree algorithm C5.0 and are evaluated with10 fold cross validation. All generated rules offer high accuracy ratings.

4. *Automatic model selection for the optimization of support vec tor machine kernels*

This approach aims to optimize the kernel parameters and to efficiently reduce the number of support vectors, so that the generalization error can be reduced drastically. The proposed methodology suggests the use of a new model selection criterion based on the estimation of the probability of error of the SVM classifier. For comparison, we considered two more model selection criteria: GACV (‘Generalized Approximate Cross-Validation’) and VC (‘Vapnik-Chernovenkis’) dimension. These criteria are algebraic estimates of upper bounds of the expected error. For the former, we also propose a new minimization scheme. The experiments conducted on a bi-class problemshow that we can adequately choose the SVM hyper-parameters using the empirical error criterion. Moreover, it turns out that the criterion produces a less complex modelwith fewer support vectors. For multi-class data, the optimization strategy is adapted tothe one-against-one data partitioning. The approach is then evaluated on images of handwritten digits from the USPS database

5. *A particle swarm optimization and pattern search based memetic algorithm forsvms parameters optimization*

Addressing the issue of SVMs parameters optimization, this study proposes an efficient memetic algorithm based on Particle Swarm Optimization algorithm (PSO) and PatternSearch (PS). In the proposed memetic algorithm, PSO is responsible for exploration of the search space and the detection of the potential regions with optimum solutions, while pattern search (PS) is used to produce an effective exploitation on the potential regions obtained by PSO. Moreover, a novel probabilistic selection strategy is proposed to select the appropriate individuals among the current population to undergo local refinement, keeping a well balance between exploration and exploitation. Experimentalresults confirm that the local refinement with PS and our proposed selection strategy are effective, and finally demonstrate effectiveness and robustness of the proposed PSO-PS based MA for SVMs parameters optimization.

6. A hyper-heuristic evolutionary algorithm for automatically designing decision-tree algorithms

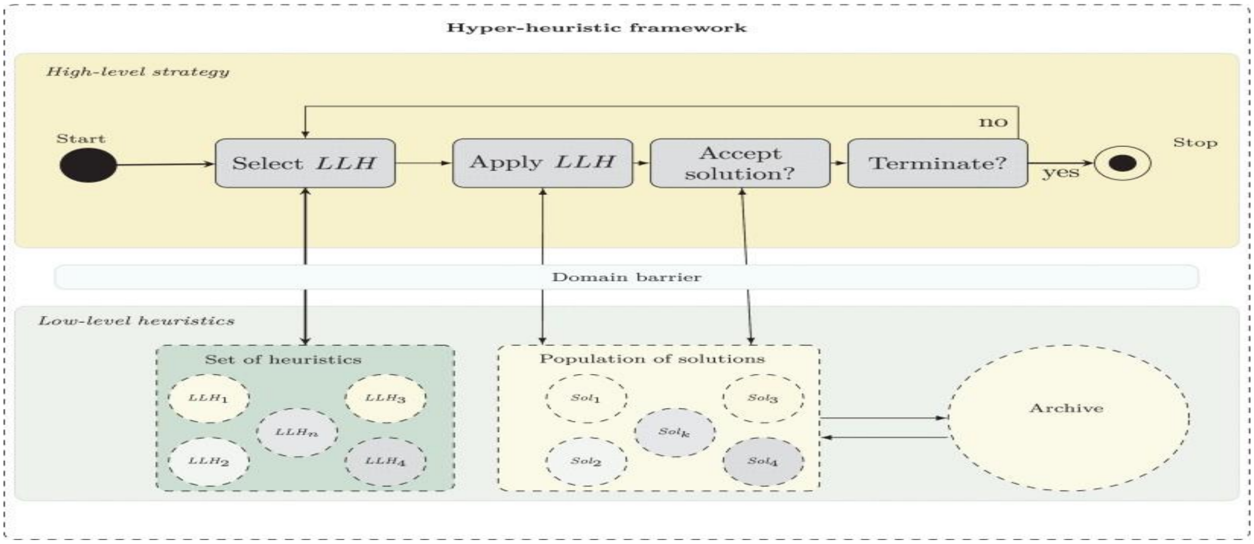
Hyper-heuristic evolutionary algorithms (HHEA) are successful methods for selecting and building new heuristics or algorithms to solve optimization or machine learning problems. They were conceived to help answer questions such as given a new classification dataset, which of the solutions already proposed in the literature is the most appropriate to solve this new problem? In this direction, we propose a HHEA to automatically build Bayesian Network Classifier (BNC) tailored to a specific dataset. BNCs are powerful classification models that can deal with missing data, uncertainty and generate interpretable models. The method receives an input a set of components already present in current BNC algorithms and a specific dataset. The HHEA then searches for the best combination of components according to the input dataset. Resultsshow the customized algorithms generated obtain results of F-measure equivalent or better than other state of the art BNC algorithms.

EXISITNG SYSTEM

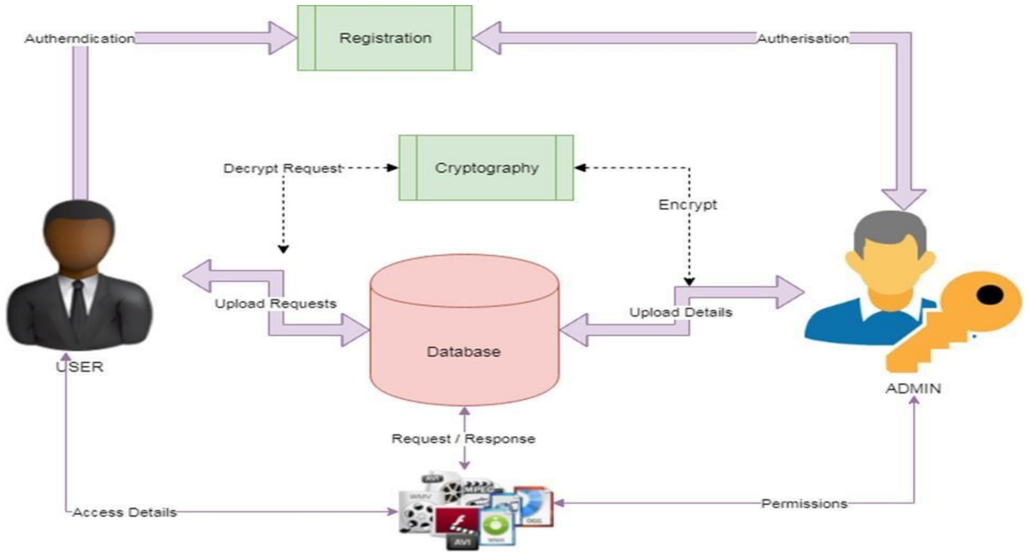
SVMs are a class of supervised learning models that have been widely used for classification and regression SVMs are based on statistical learning theory and are better able to avoid local optima than other classification algorithms. An SVM is a kernel-based learning algorithm that seeks the optimal hyper plane. The kernel learning process maps the input patterns into a higher-dimensional feature space in which linearseparation 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 nothave good generalization ability. By contrast, global kernel functions have goodgeneralization ability but a poor learning ability.

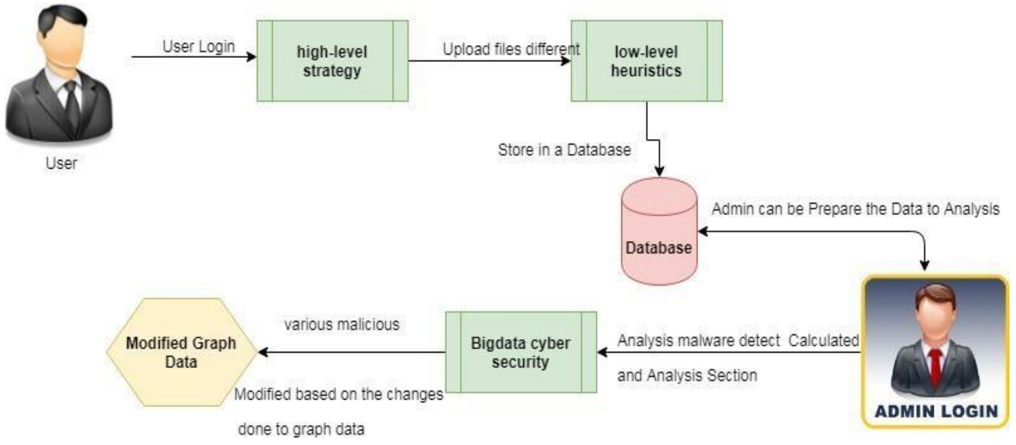
PROPOSED SYSTEM

The proposed hyper-heuristic framework for configuration selection 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 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. Thedecomposition approach operates on the population of solutions, whereas the dominance approach uses the archive. The hyper heuristic framework generates a newpopulation of solutions using the old population, the archive, or both the old populationand 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 diversityinvolves maximizing the distribution of the solutions along PF. The main components of the proposed hyper-heuristic framework are discussed in the following subsections.



SYSTEM DESIGN





IMPLEMENTATION

The project is carried out based on the following modules listed:

Approved Users

In this system users are not allowed to access resources simply. User need verify their information's with admin. Admin are the authorized and trustworthyto the network. User need to send the request to administrator that they are interested to add the community. Admin views the user request and respond withthe pass code to access users account through trusted sources like SSL (Gmail).

Security Steps and Upload

This is where the proposed algorithm is going to be effective. The admin canbe upload the files with proposed classification algorithm and cryptography in orderto classify and upload the encrypted details to network with its tag in the mark of understand to user about the resource.

Resource Access

The permissions to access the resource can be sent by users to admin. The requests have been updated by admin with the response to access the resource. Users can decrypt the resource and access the details. The important part is access the resource with the decryption. The passkey to access the details are limited. If the limit of wrong attempts over the threshold value means pass key expires.

Graphical Representation

This is graphical notation of the data given by the system. This phase of implementation will shows the effectiveness of the proposed system through pictorially in the order to better understand of proposed system.

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 ofdecomposition- and Pareto-based approaches to approximate the Pareto set of configurations.

REFERENCES

[1] M. Ahmadi, D. Ulyanov, S. Semenov, M. Trofimov, and G. Giacinto, ‘‘Novel feature extraction, selection and fusion for effective malware family classification,’’ in Proc. 6th ACM Conf. Data Appl. Secur. Privacy, 2016, pp. 183–194.

[2] A. V. Aho and M. J. Corasick, ‘‘Efficient string matching: An aid to bibliographic search,’’ Commun. ACM, vol. 18, no. 6, pp. 333–340, Jun. 1975.

[3] S. Ali and K. A. Smith-Miles, ‘‘A meta-learning approach to automatic kernel selection for support vector machines,’’ Neurocomputing, vol. 70, nos. 1–3, pp. 173–186, 2006. N.-E. Ayat, M. Cheriet, and C. Y. Suen, ‘‘Automatic model selection for the optimization of SVM kernels,’’ Pattern Recognit., vol. 38, no. 10, pp. 1733–1745, 2005.

[4] Y. Bao, Z. Hu, and T. Xiong, ‘‘A PSO and pattern search based memetic algorithm for SVMs parameters optimization,’’ Neurocomputing, vol. 117, pp. 98–106, Oct. 2013.

[5] R. C. Barros, M. P. Basgalupp, A. C. P. L. F. de Carvalho, and A. A. Freitas, ‘‘A hyper-heuristic

evolutionary algorithm for automatically designing decision-tree algorithms,'' in Proc. 14th Annu. Conf. Genet. Evol. Comput., 2012, pp. 1237–1244.

[6] M. P. Basgalupp, R. C. Barros, T. S. da Silva, and A. C. P. L. F. de Carvalho, ''Software effort prediction: A hyper-heuristic decision-tree based approach,'' in Proc. 28th Annu. ACM Symp. Appl. Comput., 2013, pp. 1109–1116.

[7] M. P. Basgalupp, R. C. Barros, and V. Podgorelec, ''Evolving decision-tree induction algorithmswith a multi-objective hyper-heuristic,'' in Proc. 30th Annu. ACM Symp. Appl. Comput., 2015, pp.110–117.

[8] A. Ben-Hur and J. Weston, ''A user's guide to support vector machines,'' in Data Mining Techniques for the Life Sciences. Methods in Molecular Biology (Methods and Protocols), O. Carugo and F. Eisenhaber, Eds. vol 609. New York, NY, USA: Humana Press, 2010, pp. 223–239.

[9] D. Brumley, C. Hartwig, Z. Liang, J. Newsome, D. Song, and H. Yin, ''Automatically identifying trigger-based behavior in malware,'' in Botnet Detection (Advances in Information Security), W. Lee, C. Wang, and D. Dagon, Eds. Boston, MA, USA: Springer, 2008.

[10] E. K. Burke, M. Hyde, G. Kendall, G. Ochoa, E. Özcan, and J. R. Woodward, ''A classification of hyper-heuristic approaches,'' in Handbook of Metaheuristics (International Series in Operations Research & Management Science), vol. 146, M. Gendreau and J. Y. Potvin, Eds. Boston, MA, USA: Springer, 2010.

[11] A. Chalimourda, B. Schölkopf, and A. J. Smola, ''Experimentally optimal ν in support vector regression for different noise models and parameter settings,'' Neural Netw., vol. 17, no. 1, pp. 127–141,

2004. C.-C. Chang and C.-J. Lin, ‘‘LIBSVM: A library for support vector machines,’’ ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, pp. 27:1–27:27, 2011.

[12] M. Chen, S. Mao, and Y. Liu, ‘‘Big data: A survey,’’ Mobile Netw. Appl., vol. 19, no. 2, pp. 171–209, Apr. 2014.

[13] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[14] M. Damshenas, A. Dehghantanha, and R. Mahmoud, ‘‘A survey on malware propagation, analysis, and detection,’’ Int. J. Cyber-Secur. Digit. Forensics, vol. 2, no. 4, pp. 10–29, 2013.

[15] K. Deb, Multi-Objective Optimization Using Evolutionary Algorithms, vol. 16. Hoboken, NJ, USA: Wiley, 2001.

[16] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, ‘‘A fast and elitist multiobjective genetic algorithm: NSGA-II,’’ IEEE Trans. Evol. Comput., vol. 6, no. 2, pp. 182–197, Apr. 2002.

[17] M. Egele, T. Scholte, E. Kirda, and C. Kruegel, ‘‘A survey on automated dynamic malware-analysis techniques and tools,’’ ACM Comput. Surv., vol. 44, no. 2, 2012, Art. no. 6.

[18] A. E. Eiben and J. E. Smith, Introduction to Evolutionary Computing, vol. 53. Heidelberg, Germany: Springer, 2003.

[19] E. Filiol, ‘‘Malware pattern scanning schemes secure against black-box analysis,’’ J. Comput. Virol., vol. 2, no. 1, pp. 35–50, 2006.

ISSN: 2057-5688

[20] E. Filiol, G. Jacob, and M. Le Liard, ‘‘Evaluation methodology and theoretical model for antiviral behavioural detection strategies,’’ J. Comput. Virol., vol. 3, no. 1, pp. 23–37, 2007.