RESEARCH ARTICLE OPEN ACCESS

A Survey on CSOAL: Cost-Sensitive Online Active Learning with Its Application to Malicious URL Detection

Miss Divyashree Kelji
divyashreekelji@gmail.com

Dr. P. D. Lambhate
jscoeit@gmail.com
SavitribaiPhule University of Pune,Department of Computer Engineering
JSPM'S JSCOE, Hadapsar, Pune

----------**********----------

ABSTRACT:

Malicious Uniform Resource Locator (URL) identification is a vital issue in web pursuit and mining, which assumes a basic job in web security. In writing, many existing examinations have endeavoured to figure the issue as a consistent administered twofold grouping undertaking, which ordinarily plans to streamline the forecast precision. Be that as it may, in a genuine malevolent URL discovery undertaking, the proportion between the quantity of malignant URLs and authentic URLs is exceptionally imbalanced, making it extremely unseemly for essentially enhancing the expectation exactness. Additionally, another key restriction of the current work is to expect a lot of preparing information is accessible, or, in other words the human marking cost could be conceivably very costly. To illuminate these issues, in this paper, so present a novel system of Cost-Sensitive Online Active Learning (CSOAL), which just inquiries a little portion of preparing information for marking and specifically streamlines two cost-delicate measures to address the class-irregularity issue. Specifically, so propose two CSOAL calculations and dissect their hypothetical execution as far as cost-delicate limits. It direct a broad arrangement of tests to look at the experimental execution of the proposed calculations for a substantial scale testing malevolent URL discovery assignment, in which the empowering results demonstrated that the proposed strategy by questioning a to a great degree little measured named information (around 0.5% out of 1-million examples) can accomplish better or exceptionally equivalent grouping execution in contrast with the best in class cost-uncaring and cost-touchy online characterization calculations utilizing an immense measure of named information.

Keywords–Uniform Resource Locator, Online Active Learning, Malicious URL.

----------**********----------

I. INTRODUCTION

The World Wide Web (WWW) allows people to access massive information on the internet, but also brings malicious information, such as fake drug, malware, and so on. A user accesses all kinds of information (benign or malicious) on the WWW b y clicking on a URL (Uniform Resource Locator) that links to a particular website. It is thus very important for internet users to evaluate the risk of clicking a URL in order to avoid accessing the malicious web sites.To tackle this challenge, researchers have attempted to investigate techniques to automatically classify whether a URL is malicious or not over the past few years, which is formally known as "malicious URL detection" [1,

2, 3, 4]. In literature, a variety of techniques have been proposed to solve the malicious URL detection problem [1, 2, 3, 4]. One major category of techniques formulates the URL detection as a classical supervised classification task and attempts to train a binary classification model in an offline learning fashion to distinguish between malicious and normal URLs. These techniques usually require collecting a considerable amount of training data in order to build a good classification model. In contrast, another category of techniques formulates it as an online supervised learning task, which is more suitable for large-scale problems. However, all these algorithms try to maximize the classification accuracy of the learnt model by assuming the ratio between the malicious and benign URLs is balanced explicitly or implicitly. It remains a very challenging research problem today, which is primarily due to several reasons. First of all, it is often a highly class-imbalanced learning problem as the number of malicious is significantly smaller than that of normal ones, which brings a critical challenge to many schemes using regular classification techniques. Second, it is usually very expensive to collect labelled data, especially the positive training data ("malicious"), which limits the application of some classical supervised classification approaches. Moreover, in a real-world application, data usually arrives in a sequential/online fashion and the size of data patterns can be very large, leading to a big challenge for developing efficient and scalable algorithms for malicious detection. To address the above challenges of malicious URL detection, in this paper, so present a novel framework of Cost Sensitive Online Active Learning (CSOAL) which can tackle malicious detection in a fairly natural, effective, and scalable approach. Unlike many existing batch learning approaches, the key idea of our framework is to formulate malicious URL detection as an online active learning task which aims to maximize the detection performance by actively querying a small amount of informative labelled data via a cost-sensitive online learning setting. In particular, so propose two CSOAL algorithms by optimizing two different cost-sensitive measures (i.e., the weighted sum of sensitivity and specificity and the weighted cost), and theoretically analyse the performance bounds of the proposed algorithms. Also further validate the empirical performance of the proposed algorithms through an extensive set of experiments for a large-scale online malicious URL detection task.

II. RELATED WORK

Web spam continues to grow in severity [1], it is imperative that the research community follow the best practices that have already been established in similar domains (e.g., email spam research). They had provided a novel method for automatically obtaining Web spam pages, and they had also presented the Webb Spam Corpus {a publicly available corpus of almost 350,000 Web spam pages that were obtained using our automated method. Webb Spam Corpus bridges the worlds of email spam and Web spam, also note that it can be used to aid traditional email spam classification algorithms through an analysis of the characteristics of the Web pages

The Webb Spam Corpus is a first-of-its-kind, large-scale, and publicly available. The interconnectivity provides preliminary support to their hypothesis. The collection process less robust against legitimate URL attacks. This is very complex system.

They propose [2] CS4VM (Cost-Sensitive Semi-Supervised Support Vector Machine) which considers unequal misclassification costs and the utilization of unlabelled data simultaneously. Experiments on a broad range of data sets show that CS4VM has encouraging performance, in terms of both the cost reduction and computational efficiency. This perception prompts an effective calculation which first gauges the mark means and afterward prepares the CS4VM with the module name.

The broad range of data sets used. Semi-supervised learning methods are usually cost insensitive. Unlabelled data are difficult to incorporate. CS4VM to multi-class scenario not possible.

They proposed [3] the Soft Confidence-Weighted (SCW) learning, a new second-order online

learning method with state-of-the-art empirical performance. The proposed SCW algorithms perform significantly better than the original CW algorithm. In this work, they extend the confidence-weighted learning for soft margin learning, which makes their Soft Confidence-Weighted (SCW) learning method more robust.
Soft Confidence Weighted (SCW) learning method more robust than the original CW.The capability to handle the non-separable cases.
The performance in terms of accuracy, number of updates, and running time cost.Performs poorly in many real-world applications.

This paper proposed [4] a novel framework of cost-sensitive online active learning (CSOAL) as a natural, simple yet fairly effective approach to tackling a real-world online malicious URL detection task. Also extensively examined their empirical performance on a large-scale real-world data set. In particular, so propose two CSOAL algorithms and analyse their theoretical performance in terms of cost-sensitive bounds. This can tackle malicious detection in a fairly natural, effective, and scalable approach.
It achieves better or highly comparable classification performance. Malicious detection in a fairly natural, effective, and scalable approach.
Information vectors to the system each one in turn and makes remedies to the system dependent on the outcomes. Specificity performance is not that much good.

Sequential decision making [5] (SDM) is an essential component of autonomous systems. Approach angle calculations have indicated significant ongoing accomplishment in settling high-dimensional successive basic leadership errands, especially in mechanical autonomy. To make agents more sample efficient, they developed a multi-task policy gradient method to learn decision making tasks consecutively, transferring knowledge between tasks to accelerate learning.
PG-ELLA gives a proficient system to online MTL of SDM undertakings while giving enhanced execution over standard approach inclination techniques.Adaptable to new changes very quickly.
The potential of PG-ELLA for cross domain is very low.

The paper proposes [6] the mechanism which supports demand response aggregator to flexibly shift the charging of electric vehicles to times where cheap but intermittent renewable energy is in high supply. Also detail a specific instance of this class, show that it achieves the optimal in realistic scenarios and demonstrate how it can be adapted to trade off efficiency with profit.
Provides online mechanisms for multi-dimensional valuations and marginal costs. It is computationally efficient and scales to hundreds of agents
Multiple units cannot be allocated to an agent in a single time step. Marginal valuations of agents are low.

This paper proposes [7] robust online multi-task learning approach in which the algorithms used are not only automatically capture the common features among all tasks and individual features for each task, but also identify the potential existence of outlier task.
Effective and efficient performance based on the closed-form updating solutions. Lower cumulative prediction error rates on both synthetic and real world datasets.
Time require to execute the algorithm's computations is high.

Second order online active learning [8], which fully exploits both first-order and second-order information to achieve high learning accuracy with low labelling cost. They conduct both theoretical analysis and empirical studies for evaluating the SOAL algorithm extensively.
Provides high learning accuracy with low labelling cost.
SOAL algorithm is able to handle extensive experiments.
The propose algorithm's performance fails for automatically re-adjusting the parameters on the learning process. Performance of SOAL in online active learning for AUC maximization is very low.

Malicious URLs [9] host unsolicited content like spam, phishing, drive-by exploits and lure unsuspecting users to become victims of scams

such as monetary loss, theft of private information, and malware installation and cause losses of billions of dollars every year. It is basic to identify and follow up on such dangers in an auspicious way. They present the formal formulation of Malicious URL Detection as a machine learning task, and categorize and review the contributions of literature studies that address different dimensions of this problem.

Provides high security from malicious URL's by using machine learning.Provides detection as a service for real-world cyber security applications.

The proposed system fails in case of smart design of closed loops system of acquiring labelled data and user feedback.

Relative Similarity Learning [10] aims to learn similarity functions from the available data with relative constraint. Online multi task relative similarity learning this method simultaneously learns multiple similarity metrics from the relative constraints data via an online learning algorithm. To further reduce the human labelling effort, they develop an active variant of OMTRSL, namely OMTRSL Active, to avoid labelling of each incoming triplet.

Provides high efficacy and efficiency in extensive experiments. Straightforward yet compelling internet learning calculation for perform multiple tasks relative comparability learning.

The proposed method is not able to handle sparse multitask learning for similarity problems. Performance of adaptive relationship matrix method of online multitask RSL is very low.

In P. Doshi et al. [11] proposes, web crawlers: their architecture, process of semantic focused crawling technology, ontology learning, pattern matching, types and various challenges being faced when search engines use the web crawlers, have been reviewed. The web results more relevant to the user query through keyword expansion have been retrieved by the system. This data is being use further for the efficient association rule.

III. PROBLEM STATEMENT

In many real applications, the dataset is usually large and unlabelled, and manually labelling all the instances is usually too expensive to afford meanwhile. To address this challenge, researchers have proposed a serial of "Online Active Learning" algorithms. Traded off URLs that are utilized for digital assaults are named as vindictive URLs. In fact, it was noted that close to one-third of all websites are potentially malicious in nature, demonstrating rampant use of malicious URLs to perpetrate cyber-crimes. A Malicious URL or a malicious web site hosts a variety of unsolicited content in the form of spam, phishing, or drive-by-exploits in order to launch attacks. In the previous experiments, the sampling factor δ was simply fixed to a constant. This experiment aims to improve the proposed CSOAL approach using the adaptive sampling factor.

IV. OBJECTIVE

Malicious Website is a general and serious issue to cyber-security. A Malicious URL or a malignant site has an assortment of spontaneous substance as spam, phishing, or drive-by-misuses with the end goal to dispatch attacks. Hence propose CSOAL approach using the adaptive sampling factor. To deal with this problem, this work classifies malicious URL based on Second Order Online Active Learning Technique.

V. SYSTEM ARCHITECTURE

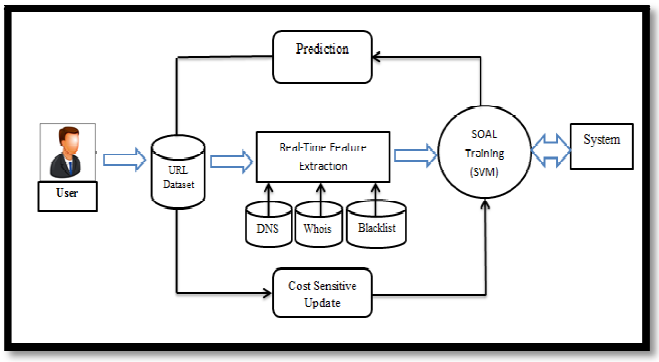

Fig. 1 System Architecture

System Architecture Explanations:

Load URL Dataset:

- In this module, a user loads URL Dataset. This dataset contains lot of URL.
- A URL has two main components: (i) protocol identifier, it indicates what protocol to use, (ii) resource name; it specifies the IP address or the domain name where the resource is located.
- The protocol identifier and the resource name are separated by a colon and two forward slashes. Traded off URLs that are utilized for digital assaults are named as malignant URLs.

Feature Extraction:

- This module extracts 3 types of features from URL Dataset.
  i. Black List Features
  ii. Lexical Features
  iii. Host-based Features
- Black List Features are a trivial technique to identify malicious URLs is to use blacklists.
- Lexical features are features obtained based on the properties of the URL name (or the URL string). The inspiration is that dependent on how the URL "looks" it ought to be conceivable to recognize pernicious nature of a URL.
- Host-based features are obtained from the host-name properties of the URL. They enable us to know the area of malevolent hosts, the personality of the noxious hosts, and the administration style and properties of these hosts.

Second Order Online Active Learning:

- This module applies Second Order Online Active Learning Training algorithm for feature extracted URL dataset.
- Online active learning has been actively explored and applied to resolve the malicious URL Detection tasks.
- In this project we used Second-order online active learning algorithm only.
- The second order online active learning aims to boost the learning efficacy by exploiting second-order information, e.g., the second order statistics of underlying distributions. For example, they usually assume the weight vector w follows a Gaussian distribution $w \sim N(\mu, \Sigma)$ with mean vector μ is belonging into Rd and covariance matrix Σ is belonging into $Rd \times d$.
- This is particularly useful for malicious URL Detection where data is sparse and high dimensional (due to the bag-of-words or alike representations of lexical features).
- This training provides some rules for predict malicious URL.
- In the previous experiments, the sampling factor δ was simply fixed to a constant. This experiment aims to examine if it is possible to further improve the proposed CSOAL approach using the adaptive sampling factor.

Prediction:

- This module predicts malicious URL based on Second Order Online Active Learning Training Results.
- It classifies a URL is malicious or benign

VI. ONLINE ACTIVE LEARNING

In this work [12], they studied online active learning in dynamic problems with potentially adversarial concept drifts. It is appeared; utilizing genuine UGC information from a news gateway at Yahoo!, that active learning is powerful for reducing labelling efforts in dynamic problems. These six models were realized by combining three active learning strategies (entropy-based criteria, the function value-based criteria, and random selection), with two bias types (unbiased and biased). Thus, in this dynamic environment, active learning provides more benefits than in the stationary analogue of the problem.

With the goal [13] of labelling the most informative instances to achieve high prediction accuracies with minimum cost, active learning is a continuously

growing area in machine learning research. A number of emerging active learning scenarios and new approaches are also discussed in this paper. Dynamic adapting just dependent on vulnerability of free and indistinguishably circulated (IID) occasions. Active learning by further taking into accounts instance correlations. In this model, Gradient Boost algorithm is used to combine a number of selector *fm*, to a strong one

A new method [14] for stopping AL based on stabilizing predictions is presented that addresses these needs. Effective methods for stopping AL are crucial for realizing the potential annotation savings enabled by AL. The proposed method is shown to fill a gap in the level of aggressiveness available for stopping AL and supports provided users with control over stopping behavior. The essential idea behind the new method is to test the predictions of the recently learned models (during AL) on examples which don't have to be labeled and stop when the predictions have stabilized.
This opens up a future zone for work on client movable ceasing.

Theoretical considerations [15] support the expectation that the combined algorithm shows superior learning speed. They conclude that second order working set selection should become the default in iterative SVM learning algorithms relying on sequential minimal optimization (SMO). Here, this calculation is significantly enhanced in speed and exactness by supplanting the working set determination in the SMO steps. This strategy requires only linear time. They conclude that second order working set selection should become the default in iterative SVM learning algorithms.

VII. MALICIOUS URL DETECTION

In this paper [16] they explored the possibility of using a confidence weighted model trained on features derived exclusively from URLs for classification. This approach differs from that of previous work they are aware of in that it does not have a dependency on any host based features. Preparing on a solitary, marked feed and testing on another totally independent feed demonstrate that this lexical component based arrangement approach is powerful. This approach uses a relatively simple feature set extracted through a simple parser. Their system is capable of detecting emerging threats as they appear and subsequently can provide increased protection against zero hour threats. Every URL is spoken to as a vector of parallel highlights that are nourished to the online calculation, in particular the certainty weighted methodology.

They use online learning algorithms [17] to detect malicious webpages. Three web based learning calculations are utilized to prepare classifiers, and their exhibitions are looked at. The heaviness of highlights is chosen by the distinction of highlight recurrence in pernicious and favourable examples. To improve the performance of online learning classifiers, an improved on line learning method is proposed. They only use URL information to determine if the URL links to malicious pages. These methods are safer, as they do not analysis the content of a webpage; they only use URL information to determine if a webpage is malicious.

The proposed [18] research work is done on streamlining the execution of the Search Engine. The proposed research tests have been directed on Shannon data gain to decide the edge estimation of dynamic dataset.

Phishing attacks [19] usually involve an attacker masquerading as a legitimate online entity to steal confidential information from the unsuspecting victims. The assailant tricks the client with social designing procedures, for example, SMS, voice, email, site and malware. They give an overview of feature extraction phase which is required by a phishing detection system as the system needs some features which can provide information about legitimacy of a website. As feature extraction is the main part in this work, they have given various feature parameters from which one can choose the desirable parameters for further classification of input website. The links provided in phishing emails draws user into entering phishing website.

VIII. CONCLUSION AND FUTURE WORK

This system proposed a novel framework of cost-sensitive online active learning (CSOAL) as a

natural, simple yet fairly effective approach to tackling a real-world online malicious URL detection task. Also presented the CSOAL algorithms to optimize cost-sensitive measures and theoretically analyse the bounds of the proposed algorithms. It extensively examined their empirical performance on a large-scale real-world data set. The encouraging results showed that (i) the proposed CSOAL method is able to considerably outperform a number of supervised cost-sensitive or cost-insensitive online learning algorithms for malicious URL detection tasks; (ii) the proposed CSOAL method is able to attain the comparable (or even better) state-of-the art predictive performance of a cost-sensitive online learner by querying a significantly small amount of labelled data (0.5% or less);and (iii) the proposed CSOAL algorithms are highly efficient and scalable for web-scale applications.

ACKNOWLEDGMENT

I profoundly grateful to Dr.P.D.Lambhatefor her expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion. I would like to express my deepest appreciation towards Dr.M.G.Jadhav Principal, Dr.S.B. Chaudhari HOD department of computer engineering and Prof. MadhavIngale PG coordinator. I must express my sincere heartfelt gratitude to all staff members of computer engineering department who helped me directly or indirectly during this course of work. Finally, I would like to thank my family and friends, for their precious support.

REFERENCES

1. S. Webb, J. Caverlee, and C. Pu, “Introducing the webb spam corpus: Using email spam to identify web spam automatically.,” in CEAS, 2006
2. Y. Li, J. T.-Y. Kwok, and Z.-H. Zhou, “Cost-sensitive semi-supervised support vector machine,” in Proceedings of the National Conference on Artificial Intelligence, vol. 1, p. 500, 2010.
3. J. Wang, P. Zhao, and S. C. Hoi, “Exact soft confidence-weighted learning,” in Proceedings of the 29th International Conference on Machine Learning (ICML-12) (J. Langford and J. Pineau, eds.), (New York, NY,USA), pp. 121–128, ACM, 2012.
4. P. Zhao and S. C. Hoi, “Cost-sensitive online active learning with application to malicious url detection,” in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’13, (New York, New York, USA), p. 919, ACM Press, 2013.
5. H. B. Ammar, U. EDU, E. Eaton, P. Ruvolo, O. EDU, M. E. Taylor, and W. EDU, “Online multi-task learning for policy gradient methods,” ICML 2014, 2014.
6. K. Hayakawa, E. H. Gerding, S. Stein, and T. Shiga, “Online mechanisms for charging electric vehicles in settings with varying marginal electricity costs,” in Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015
7. C. Zhang, P. Zhao, S. Hao, Y. C. Soh, and B. S. Lee, “Rom: A robust online multi-task learning approach,” in 2016 IEEE 16th International Conference on Data Mining (ICDM’16),, pp. 1341–1346, IEEE, 2016.
8. S. Hao, P. Zhao, J. Lu, S. C. Hoi, C. Miao, and C. Zhang, “Soal: Second order online active learning,” in Data Mining (ICDM), 2016 IEEE 16th International Conference on, pp. 931–936, IEEE, 2016.
9. D. Sahoo, C. Liu, and S. C. Hoi, “Malicious url detection using machine learning: A survey,” arXiv preprint arXiv:1701.07179, 2017.

10. S. Hao, P. Zhao, Y. Liu, S. C. H. Hoi, and C. Miao, "Online multi-task relative similarity learning," in The 26th International Joint Conference on Artificial Intelligence (IJCAI'17), 2017.
11. Poonam P. Doshi, Emmanuel M, "Web Pattern Mining Using Eclat", International Journal of Computer Applications (0975 – 8887), New York,Vol. 179 – No.8, pp.9-14, Dec2017, DOI:10.5120/ijca2017916009.
12. Wei Chu, Martin Zinkevich, Lihong Li and Achint Thomas, "Unbiased Online Active Learning in Data Streams," IEEE August 21–24, 2011.
13. Yifan Fu, Xingquan Zhu and Bin Li, "A survey on instance selection for active learning," Knowledge and Information Systems 22 July 2011.
14. Michael Bloodgood and K. Vijay-Shanker, "A Method for Stopping Active Learning Based on Stabilizing Predictionsand the Need for User-Adjustable Stopping," IEEE Natural Language Learning Sep. 2014.
15. Tobias Glasmachers and Christian Igel, "Second Order SMO Improves SVM Online and Active Learning," Volume 20 Issue 2, February 2008.
16. Aaron Blum, Brad Wardman, ThamarSolorio and Gary Warner, "Lexical Feature Based Phishing URL Detection Using Online Learning," Artificial intelligence and security, October 8, 2010.
17. Wen Zhang, Yu-Xin Ding, Yan Tang And Bin Zhao, "Malicious Web Page Detection Based On On-Line learning algorithm," International Conference On Machine Learning And Cybernetics, Guilin, 10-13 July, 2011.
18. Poonam P. Doshi, Emmanuel M, "Semantic Web Mining Using Shannon Information Gain", International Journal of Allied Practices, Research and Review, Vol 5, Issue 4, pp. 01-10, April 2018.
19. Leena&Er. AmritKaur, "Detection of Phishing Websites Using SVM Technique" IJIR of Vol-2, Issue-8, 2016.