Parameters of Genetic Algorithm with Optimization for Phishing Detection

Ms. Pallavi D. Dudhe

Prof. P.L. Ramteke

Abstract - **Phishing has become a substantial threat for internet users and a major cause of financial losses. In these attacks the cybercriminals carry out user credential information and users can fall victim. Many different anti-phishing techniques have been used to resolve phishing problem, where anti-phishing techniques are applied at both client side and server side. This paper present the detection of phishing using the genetic algorithm based approached. Genetic algorithms can be used to evolve simple rules for preventing phishing attacks. These rules are used to differentiate normal website from phishing website. These phishing websites refer to events with probability of phishing attacks. The different optimization techniques for fake website detection like Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), which can be learned in reasonable time even for large datasets and comparison of those algorithms.**

Keywords - **ant–colony, genetic, phishing, probability, swarm.**

I. INTRODUCTION

Phishing is an electronic online identity theft in which the attackers use a combination of social engineering and web site spoofing techniques to trick a user into revealing confidential information. It steals the user’s personal identity data and financial credentials [1]. Most of the phishing attacks emerge as spoofed E-Mails appearing as legitimate ones which make the users to trust and divulge into them by clicking the link provided in E-Mail.

Genetic Algorithm (GA) is a stochastic search method which has been widely used by the data mining community for discovering classification rules. The accuracy of the rules that GA finds are comparable and sometimes even more accurate than the rules obtained by the other classification algorithms. GA shows great promise in complex domains because it operates in an iterative improvement fashion. The search performed by it is probabilistically concentrated towards regions of the given data set that have been found to produce a good classification behavior [2]. In Spite of all its advantages most of the GA based classification algorithms use only a small set of training data and the task of the GA is to find out the best rule set which classifies the available instances with the lowest error rate. But today’s data generated in stock market, Super markets etc are huge and have millions of records and they do not fit to the computer memory. The Mining algorithms should be made scalable to face this challenging situation successfully. In GA scalability is mainly addressed by parallel processing or by making the solution to converge quickly and thereby reducing the number of generation which in turn reduces the learning time. Both are success full methods and extensive literature work is available for both the methods. They produce comparably accurate results and they also considerably reduce the Computational time. But several incremental learning methods have been proposed in other classification methods to address the scalability problem. They not only reduce the costs related to the accessing of the secondary storage devices but also make the classification models to adapt to the emerging new concepts. Incremental learning is an untouched area with respect to Genetic Algorithm. The proposed method in this paper tries to reduce the learning cost by incremental learning [3] [4]. It tries to build the model by examining very few records incrementally from the training data set and at the same time it also tries to maintain high accuracy level in par with other GA based classification methods which uses the whole training data set. The Genetic Algorithm is a relatively simple algorithm that can be implemented in a straightforward manner. It can be applied to a wide variety of problems including unconstrained and constrained optimization problems, nonlinear programming, stochastic programming, and combinatorial optimization problems. An advantage of the Genetic Algorithm is that it works well during global optimization especially with poorly behaved objective functions such as those that are discontinuous or with many local minima.

There is an efficient model which is based on using Association and classification Data Mining algorithms optimizing with PSO algorithm. These algorithms were used to characterize and identify all the
factors and rules in order to classify the phishing website and the relationship that correlate them with each other. It also used MCAR classification algorithm to extract the phishing training data sets criteria to classify their legitimacy. After classification, those results have been optimized with Ant Colony Optimization (ACO) algorithm.

II. LITERATURE REVIEW

Traditional classification methods like C4.5 are designed to deal with static data. Subsequently many new non GA based methods have been proposed to deal with concept drifts [6]. An incremental GA was proposed by Gaun et al which updates the rules based on the new data. Due to the arrival of new data or new attribute or class, the classification model may change. So to deal with this the author proposes an incremental based GA. Huai li et al proposes a memory based incremental genetic algorithm to deal with the concept drift. They make an assumption that the new training data pass through a fixed-size window at a steady rate. So now a days most faster and quiker genetic algorithm will be area of work.

III. Heuristic-Based Approach

Heuristic-based anti-phishing technique is to estimate whether a page has some phishing heuristics characteristics. For example, some heuristics characteristics used by the Spoof Guard toolbar include checking the machine name, checking the URL for available common spoofing techniques, and checking against previously seen images [1] [10]. If we only use the Heuristic-based technique, the accuracy is not enough. Its pages are often similar with the legitimate sites. Therefore, some researchers proposed a similarity assessment method to detect phishing sites. It provides an efficient checking mechanism where hostname, URL, images are checked to detect phishing.

IV. Proposed Method

In the proposed method the scalability issue is addressed in a different perspective with respect to GA. In classification problems GA should use all the records in the training set to calculate the potential solutions fitness values and for very large data sets this constitutes the most part of the learning cost [4]. By minimizing the number of records to be examined proportionate learning cost can be saved. The core idea of the paper is to make GA to learn incrementally and because of it the number of records needed to build an accurate model can be reduced
Considerably [5]. To reduce the learning time with respect to data sets that do not fit to the main memory, normal procedure followed by the data mining community is to partition the data set and mine incrementally.

4.1 Genetic Algorithm

Genetic algorithms can be used to evolve simple rules for preventing phishing attacks. These rules are used to differentiate normal website from phishing website [8]. These phishing websites refer to events with probability of phishing attacks. The rules stored in the rule base are usually in the following form :

if { condition } then { act }

For example, a rule can be defined as:

If { The IP address of the URL in the received e-mail
Finds any match in the Rule set}
Then
{Phishing e-mail
}

This rule can be explained as: if there exists an IP address of the URL in e-mail and it does not match the defined Rule Set for White List then the received mail is a phishing mail.

It provides the feature of malicious status notification before the user reads the mail. It also provides malicious web link detection in addition of phishing detection.

In the genetic algorithm process is as follows:

Step 1: Determine the number of chromosomes, generation, and mutation rate and crossover rate value.

Step 2: Generate chromosome-chromosome number of the population, and the initialization value of the genes chromosome-chromosome with a random value.

Step 3: Process steps 4-7 until the number of generations is met.

Step 4: Evaluation of fitness value of chromosomes by calculating objective function.

Step 5: Chromosomes selection.

Step 6: Crossover.

Step 7: Mutation.

Step 8: New Chromosomes (Offspring).

Step 9: Solution (Best Chromosomes).

The figure 1 shows the flow of genetic algorithm step by step. The parameters of genetic algorithm are as follows:

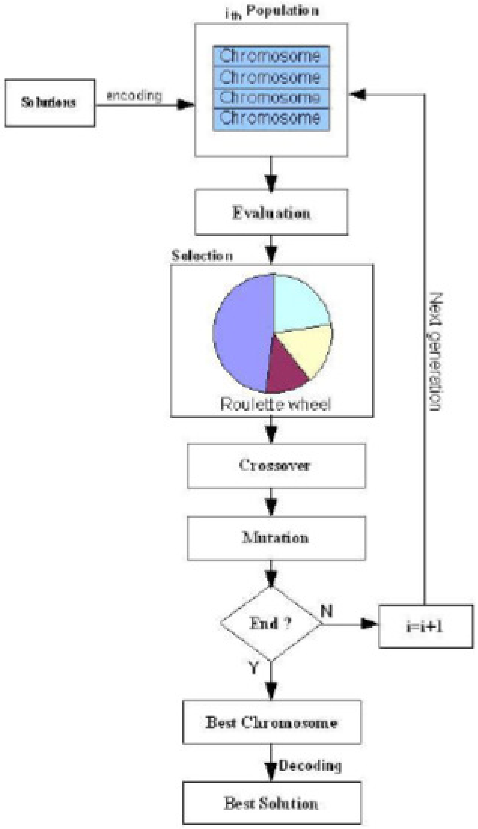

Figure 1: Flow of genetic algorithm

4.1.1 Chromosome representation

In GA applications, the unknown parameters are encoded in the form of strings, so-called chromosomes. A chromosome is encoded with binary, integer or real numbers. Since multi-spectral image data are usually represented by positive integers, in this research a chromosome is encoded with units (tuples) of positive integer numbers. Each unit represents a combination of brightness values, one for each band, and thus a potential cluster centroid. The length of the chromosome, *K*, is equivalent to the number of units and thus of potential clusters in the classification problem. *K* is selected from the range [*Kmin, Kmax*], where *Kmin* is usually assigned to 2 unless special cases are considered. and *Kmax* describes the maximum chromosome length, which means the maximum number of possible cluster centroids. *Kmax* must be selected according to experience [9]. Without assigning the number of clusters in advance, a variable string length is used. Invalid (non-existing) clusters are represented with negative integer "-1". The values of the different chromosomes are then changed in an iterative process involving different rules (called crossover and mutation) to determine the correct number of clusters (the number of valid units in the chromosomes) and the cluster centroids for a given Classification problem.

4.1.2 Chromosome initialization

A population is the set of chromosomes. The typical population size can range from 20 to 1000 . The following example is given to explain the representation a the population: we assume to have a satellite image with three bands; *Kmin* is set to 2 and *Kmax* to 8. At the beginning, for each chromosome *i* (*i =1, 2,…,.P,* where *P* is the size of population) all values are chosen randomly from the data space (universal data set; here: positive integers with the appropriate radiometric resolution). Such a chromosome belongs to the so-called parent generation. One (arbitrary) chromosome of the parent generation is given here (note that it contains only five valid centroids, since “-1” appears three times in this chromosome):

-1 (110,88,246) (150,78,226) -1 (11,104,8) (50,100,114) – 1 (227,250 192)

4.1.3 Selection and crossover

The purpose of selection and crossover (the latter is also called recombination) is to create two new individual chromosomes from two existing chromosomes selected randomly from the crossover pool. The crossover pool contains a percentage (the so called crossover percentage) of the current population, which constitutes the best chromosomes according to the chosen index [10]. Typical crossover operations are one-point crossover, two-point crossover, cycle crossover and uniform crossover.

4.1.4 Mutation

Mutation follows crossover. During mutation, all the chromosomes in the population are checked unit by unit and according to a pre-defined probability all values of a specific unit may be randomly changed.

4.1.5 The Fitness Function (Index)

Based on crossover and mutation, the chromosomes, once initiated, iteratively evolve from one generation to the next. In each generation the fitness function (index) is used to measure the fitness or adaptability of each chromosome in the population. After calculating the index for each chromosome of a given population, the best chromosome is compared to the best one of the previous generation (iteration). The termination condition for the iterations is that the difference between these two values lies below a pre-defined threshold. In case this condition holds the best chromosome of the current generation is considered as the final result, it contains the number of clusters (number of units with values different from “-1”) and the cluster centroids (the values of the valid units). If the termination condition is not met, the best chromosomes are selected into the crossover pool and after mutation a new iteration is started [11]. The population thus evolves over generations in the attempt to maximize or stabilize the index. The computations are also stopped once a maximum number of generations is reached.

V. Optimization Techniques

The goal of this research is to investigate the suitability of different optimization techniques for fake website detection like Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), which can be learned in reasonable time even for large datasets and comparison of those algorithms.

5.1 Optimization

Optimization is nothing but selection of a best element from some set of available deternatives. An optimization problem consists of maximizing or minimizing a real function by systematically choosing input values from within an allowed set and computing the value of the function.

5.2 Ant Colony Optimization (ACO)

Our proposed genetic algorithm uses ACO algorithm for finding interesting relationships among data items. It uses its evolutionary capability to efficiently find more interesting subsets of association rules. It does not exhaustively search for all possible association rules as conventional ARM approaches does [12] . In each generation of the algorithm a number of rules that satisfies minimum support and confidence threshold are selected for the final classifier. After each generation pheromones values are updated in such a way that better rules can be extracted in next coming generations [13]. The final discovered rule set is the predictive model and is used to classify unseen test samples. In this work, a hybrid classification algorithm called ACO combining the idea of association rules mining and supervised. The proposed technique integrates classification with association rule mining to discover high quality rules for improving the performance of resulting classifier. ACO is used to mine only appropriate subset of class association rules instead of exhaustively searching for all possible rules. The mining process stops when the discovered rule set achieves a minimum coverage threshold.

5.3 Particle Swarm Optimization (PSO)

A particle is treated as a point in an M-dimensional search space, and the status of a particle is characterized by its position and velocity. Initialized with a swarm of random particles, PSO is achieved through particle flying along the trajectory that will be adjusted based on the best experience or position of the one particle called local best and the best experience or position ever found by all particles called global best. The M-dimensional position for the ith particle in the tth iteration can be denoted as

$$X\ i(t) =\{\ x\ i1(t),\ xi2(t),\ldots.xiM(t)\}$$

Similarly, the velocity also a multi dimensional vector, fo the ith particle in the tth iteration can be described as

$$V\ i(t) =\{\ v\ i1(t),\ vi2(t),\ldots.viM(t)\}$$

Particle swarm has two primary operators : Velocity update and Position Update given by

$$V\ i(t) =w(t)\ V\ i(t-1)+c1r1(XiL-Xi(t-1)+c2r2(XiG-Xi(t-1))$$

$$Xi(t) =Xi(t-1) + V\ i(t)$$

where XiL represents the local best of the ith particle and XG represents the global best of the ith iteration, c1 and c2 are positive constants and r1 and r2 are random numbers between 0 to 1 and w(t) is the inertia weight used to control the impact of the previous velocities on the current velocity, influencing the tradeoff between the global and local experiences[8]. In this paper PSO technique has been applied to get optimal solution to reduce the optimization time. It is also used to accelerate the search for the domain block that is most similar to the range block [14]. Here the particles represent the solutions. Initially the particles take some random positions in solutions. Most of evolutionary techniques have the following procedure:

1. Random generation of an initial population
2. Reckoning of a fitness value for each subject. It will directly depend on the distance to the optimum.
3. Reproduction of the population based on fitness values.
4. If requirements are met, then stop. Otherwise go back to 2.

From the procedure, we can learn that PSO shares many common points with GA. Both algorithms start with a group of a randomly generated population, both have fitness values to evaluate the population. Both update the population and search for the optimum with random techniques. Particles update themselves with the internal velocity. They also have memory, which is important to the algorithm. Compared with ACO, the information sharing mechanism in PSO is significantly different. In PSO, only gBest (or lBest) gives out the information to others. It is a one -way information sharing mechanism. The evolution only looks for the best solution. Compared with ACO, all the particles tend to converge to the best solution quickly even in the local version in most cases.

VI. Conclusion

In the above study we can conclude that most of anti-phishing techniques rely on content of web sites, URL to detect phishing. To make the Genetic Algorithm applicable for generating rules for large data sets it should be made scalable. Genetic Algorithm has a number of free parameters. Two of them, namely population size and the crossover probability were considered in this research. In our results the population size proofed to be significantly more important than the crossover probability. In future research we will further investigate the potential influence of the other parameters and also consolidate our results using more test data and alternative indices for measuring the chromosome fitness.

References

[1] L. A. T. Nguyen, B. L. To, H. K. Nguyen, and M. H. Nguyen, "Detecting phishing web sites: A heuristic URL-based approach," in Advanced Technologies for ommunications (ATC),International Conference on, pp. 597- 602,2013.

[2] Linyu Yang, Dwi H. Widyantoro, Thomas Ioerger and John Yen, “An Entropy-based Adaptive Genetic Algorithm for Learning Classification Rules”, *Proceedings of the 2001 Congress on Evolutionary Computation,* , Issue , Page(s):790 – 796,2001.

[3] Bandyopadhyay, S., and U., Maulik, Genetic clustering for automatic evolution of clusters and application to image classification, *IEEE Pattern Recognition*, 35:1197-1208,2002.

[4] Rothlauf, F., *Representations for Genetic and Evolutionary Algorithms*, Springer, Netherlands, 314p, 2006.

[5] J.shreeram, M.subam, P.shanthi, K.manjula. Sastra University Kumbakanam “Anti phishing detection of phishing attacks using genetic algorithm”.Retrieved on October 8, 2010.

[6] I-hui li, I-en liao and Wei-zhi pang, “Mining classification rules in the presence of Concept drift with an incremental genetic Algorithm “, *journal of theoretical and applied information technology,* 2008.

[7] Jing Gao, Bolin Ding, Wei Fan, Jiawei Han,Philip S.Yu, “Classifying Data Streams with Skewed Class Distributions and Concept Drifts”**,** *IEEE Internet Computing, Special Issue on Data Stream Management*(IEEEIC),Nov/Dec., page(s)37-49, 2008.

[8] T. Venkat Narayana Rao et al., “ / Genetic Algorithms and Programming-An Evolutionary Methodology”, *International Journal of Computer Science and Information Technologies*, Vol. 1 (5), 427-437, 2010.

[9] V. Shreeram, etc., “Anti-phishing detection of phishing attacks using genetic algorithm”, Proceedings of ICCCCT’10, pp. 447-450, 2010

[10] Sophie Gastellier-Prevost, etc., “Decisive heuristics to differentiate legitimate from phishing sites”, *Proceedings of IEEE*, 2011.

[11] Poonam Garg “A Comparison between Memetic algorithm and Genetic algorithm for the cryptanalysis of Simplified Data Encryption Standard algorithm”, International Journal of Network Security & Its Applications (IJNSA), Vol.1, No 1, April 2009

[12] Xian-Jun Shi Hong Lei , “ A Genetic Algorithm-Based Approach for Classification Rule Discovery”. *International Conference on Information Management, Innovation Management and Industrial Engineering, 2008***,** Volume: 1, page(s): 175-178, 2008.

[13] A. Hossain, M. Dorigo, Ant colony optimization web page, http:// iridia.ulb.ac.be / mdorigo/ACO/ACO.html N. Ascheuer, Hamiltonian path problems,2002.

[14] Alaa Aljanaby, Ku Ruhana Ku Mahamud, Optimizing Large Scale Combinational Problems Using Multiple Ant Colonies Algorithm based on Pheromone Evaluation technique,2004.

Author Profile

Miss Pallavi D. Dudhe has Completed Degree in B. Tech(Computer Science and Engineering) from Government College of Engineering, Amravati, Maharashtra, India. She is pursuing M.E. Computer Science and Engineering in HVPM COET, Amravati, Maharashtra, India.

Prof. P. L. Ramteke is working as an Associate Professor and Head of Department (HOD) Department of Information Technology, HVPM COET, Amravati, Maharashtra, India