Volume 4, Issue 3, March 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey on Data Mining Networks Using IDS with SVMS and Ant Colony Networks

Suraj Singh Tomer, Vivek Kumar Singh
Department of Computer Science and Engineering
LNCTS, Bhopal, India

Abstract— Data is core of any organization so competitors trying to get the critical information from that data to find out the way to get contracts or projects. Some of them hire hackers to steal company valuable data. Hacker uses software tools and own knowledge to find vulnerability in the applications or network. Now it is very important to protect data from such theft or attacks by hackers. Intrusion detection system (IDS) can analyzes the system and network activity for unauthorized entry and malicious activity. We introduce machine learning based data classification algorithm and approach in intrusion detection by combining modified SVM and CSOACN. The IDS based on the new algorithm can be applied as pure SVM, pure CSOACN and combination of modified SVM and CSOACN by constructing the detection classifier different training modes. The combined SVM and CSOACN will reduces the update time of training sample and also capable to identified new malicious behaviour of user.

Keywords—Data mining; Data classification; Intrusion Detection System; IDS; Machine learning; Support vector machine; SVM; Ant colony optimization; CSOACN.

I. INTRODUCTION

Intrusion detection is the detecting of actions that attempt to compromise the integrity, confidentiality or availability of a resource in the network. In case of an intrusion, an IDS (Intrusion Detection System) detects it quickly and takes appropriate action. The combination of facts, such as the rapid growth of the Internet, the huge financial possibilities opening up in electronic trade, and the lack of truly secure systems, makes IDSs an important front edge research orientation of network security. Although many different IDSs have been developed, the detection schemes generally fall into one of the two categories: anomaly detection or misuse detection.

Anomaly detectors look for behavior that deviates from normal use of system whereas misuse detectors look for behavior that matches a known attack scenario. The normal behavior is based on lots of innocuous factors and highly variable. Therefore, the selection of features to monitor is the main issue in anomaly detection. The approach of misuse detection is to model the abnormal system behaviors at first and define any other behavior as normal behavior. Namely, known intrusion attacks are represented in the form of pattern or signature, activities that match those attack scenarios that can be detected and different suitable actions will be further taken for various intrusions. The main issue in misuse detection systems is the pattern recognition and signature depiction of the pertinent attack, which should encompass all possible variations but should not match non-intrusive activities.

A. The Data Mining Task

The data mining tasks are of various sorts counting on the utilization of knowledge mining result the information mining tasks are classified as:

a) *Exploratory Data Analysis*

In the repositories huge amount of information is available. This data mining task will serve without the knowledge for what the customer is searching then it analyzes the data. These techniques are interactive and visual to the customer.

b) *Descriptive Modelling*

It describe all the data, it includes models for overall probability distribution of the data, partitioning of the p dimensional space into groups and models describing the relationships between the variables.

c) *Predictive Modelling*

This model permits the value of one variable to be predicted from the known values of other variables.

d) *Discovering Patterns and Rules*

This task is primarily used to find the hidden pattern as well as to discover the pattern in the cluster. In a cluster , a number of patterns of different sizes are available .The aim of this task is "how best we will detect the patterns" .This can be accomplished by using rule induction and other techniques in the data mining algorithm like(K-Means / K-Medoids) .

B. *Retrieval by Content*

The primary objective of this task is to find the data sets, of frequently used in the audio/video as well as images. It is finding pattern similar to the pattern of interest in the data set.

C. Types of Data Mining System

Data mining systems can be categorized to various forms the classification is as follows:

a) Classification of data mining systems according to the type of data source mined

In an organization a huge amount of data is available where we need to classify these data but these are available most of times in a similar fashion. We need to classify these data according to its type (maybe audio/video, text format etc).

b) *Classification of data mining systems according to the data model*

There are various data mining models (Relational data model, Object Model, Object Oriented data Model, Hierarchical data Model/W data model) are available, In each and every model is using different data .According to these data model the data mining system classify the data in the model.

c) *Classification of data mining systems according to the kind of knowledge discovered*

This classification is based on the kind of knowledge discovered or data mining functionalities, such as classification, inequity, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several data mining functionalities together.

d) *Categorization of data mining systems according to mining techniques used*

This classification is according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented, etc.

The classification can also take into account the degree of user interaction involved in the data mining process such as query-driven systems, interactive tentative systems, or autonomous systems. A comprehensive system would provide a wide variety of data mining techniques to fit different situations and options, and offer different degrees of user interaction.

D. Support Vector Machine

Support Vector Machine (SVM), is one of best machine learning algorithms, which was proposed in 1990's and used mostly for pattern recognition. This has also been applied to many pattern classification problems such as image recognition, speech recognition, text categorization, face detection and faulty card detection, etc. Pattern recognition aims to classify data based on either a priori knowledge or statistical information extracted from raw data, which is a powerful tool in data separation in many disciplines. SVM could be a supervised sort of machine learning. rule during which, given a group of coaching examples, every marked as happiness to one of the numerous classes, associate degree SVM coaching rule builds a model that predicts the class of the new example. SVM has the bigger ability to generalize the matter that is that the goal in applied math learning [15].

E. Ant Colony Optimization

In the early 1990s, ant colony optimization (ACO) was introduced by M. Dorigo and colleagues as a novel nature-inspired met heuristic for the solution of hard combinatorial optimization (CO) problems [6]. ACO belongs to the class of met heuristics, which are approximate algorithms used to obtain good enough solutions to hard CO problems in a reasonable amount of computation time. Other examples of met heuristics are taboo search, simulated annealing, and evolutionary computation. The inspiring source of ACO is the foraging behavior of real ants. When searching for food, ants initially explore the area surrounding their nest in a random manner.

As presently as associate degree hymenopter on finds a food source, it evaluates the amount and therefore the quality of the food and carries a number of it back to the nest, throughout the comeback trip, the hymenopter on deposits a chemical secretion path on the bottom. The amount of secretion deposited, which can rely upon the amount and quality of the food, can guide different ants to the food supply. This characteristic of real hymenopter on colonies is exploited in artificial hymenopter on colonies so as to unravel CO issues.

II. **RELATED WORK**

In 2013 by Wenying Feng et al proposed a new algorithm (CSVAC) for generating classifiers with clustering, and applied it to the intrusion detection problem. This approach combines two existing machine learning methods (SVM and CSOACN) to achieve better performance in both detection accuracy rate and faster running time. The basic task is to classify network activities (in the network log as connection records) as normal or abnormal while minimizing the classification. Although different classification models have been developed for network intrusion detection, each of them has its strengths and weaknesses, including the most commonly applied Support Vector Machine (SVM) method and the Clustering based on Self-Organized Ant Colony Network (CSOACN). This new approach combines the SVM method with CSOACNs to take the advantage of both while avoiding their weaknesses. Their proposed algorithm is implemented and evaluated using a standard benchmark KDD99 data set. Experiments show that CSVAC (Combining Support Vectors with Ant Colony) outperforms SVM alone or CSOACN alone in terms of both classification rate and run-time efficiency [1].

Here author explained the need to apply data mining methods to network events to classify network attack events. They summarized the results of earlier studies and explored the earlier models on the performance improvement of the Naïve Bayes model in data mining and introduced the HNB model as a solution to the intrusion detection problem. They augmented the Naïve Bayes and structurally extended Naïve Bayes methods with the leading discretization and feature selection methods to increase the accuracy and decrease the resource requirements of intrusion detection problem. They compared the performance of the Naïve Bayes and the leading extended Naïve Bayes approaches with the new HNB approaches of an intrusion detection system. The results of our experimental study, which uses the KDD'99 dataset, show that the Hidden Naïve Bayes multiclass classification model augmented with various discretization and feature selection methods exhibits better overall results in terms of detection accuracy, error rate and misclassification cost than the traditional Naïve Bayes model, the leading extended Naïve Bayes models and the KDD'99 winner. The results also

indicate that their model significantly improves the detection of denial-of-service attacks compared with the other models [2].

The Reda M. et al. discusses the data-mining-based network intrusion detection systems. Data-mining techniques are utilized in misuse, anomaly, and hybrid detection. First, the random forests rules employee knowledge mining classification rule into a misuse detection methodology to create intrusion patterns from a balanced coaching dataset, and to classify the captured network connections to the most forms of intrusions to the designed patterns. The most downside of the misuse notice on methodologies that it cannot detect novel intrusions that aren't trained on before. Secondly, the k-means rule is employed as a data-mining clump rule into supervised anomaly detection methodology to partition the captured network connections into a nominative variety of clusters, and then notice the abnormal clusters reckoning on their options.

The "KMlocal" implementation of the k-means clump rules employed to implement our anomaly detection methodology. And there methodology is evaluated over the KDD'99 data sets on solution the issues of categorical and completely different scales options. The most downside of the anomaly detection methodology is that the high false positive rate. Thirdly, the random forest rules employed with the wk means rule to create a hybrid frame work to beat the drawbacks of each misuse and anomaly detection. Feature importance values calculated by the random forests rule are utilized in the misuse detection half to boost the detection rate of the anomaly detection half. A supervised methodology is planned to boost normal cluster determination by injecting glorious attacks into the unsure information before being clustered, and victimization these glorious intrusions in deciding the ab normal clusters [3].

Qinglei Zhang and Wenying Feng present a framework for a new approach in intrusion detection by combining two existing machine learning methods (i.e. SVM and CSOACN). The IDS based on the new algorithm can be applied as pure SVM, pure CSOACN or their combination by constructing the detection classifier under three different training modes respectively. The initial experiments indicate that performance of their combination is better than pure SVM in terms of higher average detection rate as well as lower rates of both negative and positive false and is better than pure CSOACN in terms of less training time with comparable detection rate and false alarm rates. The system can be used in different cases by constructing the detection classifier under three different training modes (i.e. SVM mode, CSOACN mode, CSVAC mode). SVM training mode is suitable for the time intense case that only one binary classifier is required by training upon a small amount of labeled data. Namely, it only needs to distinguish the data of normal from abnormal. On the other hand, the CSOACN training mode is suitable for the preciseness intensive case and can solve multiclass problems upon both label and unlabeled data. The CSVAC mode, which is based on the combination of SVM and CSOACN, can be used to balance the performance of IDS in terms of efficiency and accuracy [4].

In 2012 by Ashis Pradhan gives the concept about support vector machine (SVM) is one of the most important machine learning algorithms that has been implemented mostly in pattern recognition problem, for e.g. classifying the network traffic and also in image processing for recognition. Lots of research is going on in this technique for the improvement of QOS (quality of service) and in security perspective. The latest works in this field have proved that SVM performs better than other network traffic classifier in terms of generalization of problem.

SVM based machine learning technique has been implemented mostly in pattern recognition in networks field as this technique has proven to be efficient than the any other most used pattern recognition technique, for e.g. Neural Network, Basiean classification. The performance given by the SVM is comparatively higher if it involves large dataset for generalization of problem. The major strength of SVM is that the training of data is relatively easy.

Applications: Since, Support Vector Machine is supervised machine learning algorithm which performs on training basis, so it has mostly been implemented in network areas. For example: classifying the different network application like FTP, HTTP, P2P, etc. The other works of SVM are:-Text classification, Speech recognition, Image clustering for image compression and also Image classification, hand written digit recognition problem, and many other application that requires pattern recognition technique. The SVM can also be implemented in BOTNET detection for isolation of malicious traffic, for improvement in network traffic security. Also some works can be implemented using SVM by filtering network traffic to enhance QOS (Quality of Service). The latest works in this algorithm have proven it that it can be used in recognition of shape and hand gesture in static and also in dynamic environment [5].

Here author over-viewed some recent efforts to develop a theory of ant colony optimization. After giving a brief introduction to the algorithms and problems considered in the overview, they have discussed convergence, presented connections between ACO algorithms and the stochastic gradient ascent and cross-entropy methods within the framework of model-based search, and finally discussed the influence of search bias on the working of ACO algorithms.

Research on a new met heuristic for optimization is often initially focused on proof of concept applications. It is only after experimental work has shown the practical interest of the method that researchers try to extend their understanding of the method's functioning not only through more and more sophisticated experiments but also by means of an effort to build a theory. Tackling questions such as "how and why the method works'' is important, because finding an answer may help in improving its applicability.

Ant colony optimization, which was introduced as a novel technique for solving hard combinatorial optimization problems, finds itself currently at this point of its life cycle. With this article they provide a survey on theoretical results on ant colony optimization. First, we review some convergence results. Then they discuss relations between ant colony optimization algorithms and other approximate methods for optimization. Finally, they focus on some research efforts directed at gaining a deeper understanding of the behavior of ant colony optimization algorithms. Throughout this concept they identify some open questions with a certain interest of being solved in the near future [6].

In 2013 by Meng Lingxi et al. gives the concept about, Intrusion detection is an important aspect of the network information safety. For the disadvantage that the existing intrusion detection method is not comprehensive of various kinds of attack and has lower detection rate and the higher fault detection rate, an improved ant colony clustering method for intrusion detection is proposed. The convergence rate of ant colony cluster algorithm is improved. In the optimization process, the information entropy is introduced to prevent the local optimal, and thus the method can adjust the automatic the update pheromone and improve the clustering speed. The experimental results show that the method not only improves the detection rate, but reduced the fault detection rate, and can detection precisely the various kinds of attacks [7].

Fernando E. B. Otero et al. proposes a new sequential covering strategy for ACO classification algorithms to mitigate the problem of rule interaction, where the order of the rules is implicitly encoded as pheromone values and the search is guided by the quality of a candidate list of rules. His experiments using 18 publicly available data sets show that the predictive accuracy obtained by a new ACO classification algorithm implementing the proposed sequential covering strategy is statistically significantly higher than the predictive accuracy of state of the art rule induction classification algorithms.

Ant colony optimization (ACO) algorithms have been successfully applied to discover a list of classification rules. In general, these algorithms follow a sequential covering strategy, where a single rule is discovered at iteration of the algorithm in order to build a list of rules. The sequential covering strategy has the drawback of not coping with the problem of rule interaction, i.e., the outcome of a rule affects the rules that can be discovered subsequently since the search space is modified due to the removal of examples covered by previous rules [8].

Here author proposed a rapid and accurate machine learning approach which is developed to predict the winding ac resistance of air-core reactors. By applying the pairing comparison method to the finite-element simulations of real reactor models, reliable and simplified models are derived by eliminating the factors that have a negligible influence on the winding ac resistance. The support vector machine (SVM) approach is introduced into building a regressive function for calculating the ac resistance of layered windings. In the SVM-based learning algorithm, a 3-degree resistance factor kernel is proposed through factorial experiment and kernel construction [9].

Ant colony algorithm may produce redundant states in the graph, it's better to minimize such graphs to enhance the behavior of the inducted system. By moving, each ant incrementally constructs a solution to the problem. . When an ant complete solution, or during the construction phase, the ant evaluates the solution and modifies the trail value on the components used in its solution. Ants deposit a certain amount of pheromone on the components; that is, either on the vertices or on the edges that they traverse. The amount of pheromone deposited may depend on the quality of the solution found. Subsequent ants use the pheromone information as a guide toward promising regions of the search space. Ants adaptively modify the way the problem is represented and perceived by other ants, but they are not adaptive themselves. The genetic programming paradigm permits the evolution of computer programs which can perform alternative computations conditioned on the outcome of intermediate calculations, which can perform computations on variables of many different types, which can perform iterations and recursions to achieve the desired result, which can define and subsequently use computed values and sub-programs, and whose size, shape, and complexity is not specified in advance [10].

This paper presents an approach for solving traveling salesman problem based on improved ant colony algorithm. The main contribution of this paper is a study of the avoidance of stagnation behavior and premature convergence by using distribution strategy of initial ants and dynamic heuristic parameter updating based on entropy. Then a mergence of local search solution is provided. The experimental results and performance comparison showed that the proposed system reaches the better search performance over ACO algorithms do [11].

The purpose of author for these concepts is to apply ACO approach to the portfolio optimization mean variance model. The problem of portfolio optimization is a multi objective problem that aims at simultaneously maximizing the expected return of the portfolio and minimizing portfolio risk. Present study is a heuristic approach to portfolio optimization problem using Ant Colony Optimization technique. The performance of ACO is compared with front on function of MATLAB software as an exact method. The results show that proposed ACO approach is reliable but not preferred to an exact method [12].

Genetic Algorithms (GA) have been used to evolve computer programs for specific tasks, and to design other computational structures. The recent resurgence of interest in AP with GA has been spurred by the work on Genetic Programming (GP). GP paradigm provides a way to do program induction by searching the space of possible computer programs for an individual computer program that is highly fit in solving or approximately solving the problem at hand[13].

The traditional text-based image classification fails to recognize the underlying content of image that would lead to misclassifying issue. In this paper, a new classification based on weighted Euclidean distance whereas the weights are estimated via Support Vector Machine (SVM) is proposed. To overcome the problem of misclassification and increase the classifier accuracy for some particular classes, the new classification method uses the weight set estimated from SVM, which is then applied to the Euclidean-based K-Nearest Neighbor (K-NN) method. The experiments have revealed the proposed method has an accuracy of 93.75%, which is better than traditional classification such as K-NN and SVM. Moreover, the accuracy of the image feature\ vector has minimal impact on the accuracy of the method compare to other classifier. The SVM weighted K-NN classifier has provided a promising direction for criteria based image retrieval [14].

III. PROBLEM STATEMENT

In pure SVM there is some room for optimization, they are as follows

- Optimize binary SVM classification rules,
- Train conventional linear classification SVMs optimizing error rate in time that is linear in the size of the training data through an option, but corresponding formulation of the instruction problem [9].

This could be much faster than pure SVM for large training sets. In the previous IDS pure SVM is used, if we will use modified SVM with above mention point can improve overall performance of IDS.

IV. CONCLUSION & FUTURE WORK

We have reviewed various Intruder Detection System approaches and combination of data mining techniques to enhance the detection of malicious activities in the network. The data classification technique SVM is widely used in Intruder Detection System. We can enhance the detection rate and minimize the training time by using modified SVM with the combination of CSOACN. Optimized ant colony network finds new behaviour of network users. New behaviour is now classified into malicious or normal by the classifier. The combination of data mining techniques improves the detection rate of IDS.

REFERENCES

[1] Wenying Feng, Qinglei Zhang, Gongzhu Hu and Jimmy Xiangji Huang, "Mining network data for intrusion detection through combining SVMs with ant colony networks", Future Generation Computer Systems, 2013, in Press.

[2] Levent Koc, Thomas A. Mazzuchi and Shahram Sarkani, "A network intrusion detection system based on a Hidden Naïve Bayes multiclass classifier", Expert Systems with Applications, Vol. 39, Issue 18, , pp 13492–13500, December 2012.

[3] Reda M. Elbasiony, Elsayed A. Sallam, Tarek E. Eltobely and Mahmoud M. Fahmy, "A hybrid network intrusion detection framework based on random forests and weighted k-means", Ain Shams Engineering Journal, Vol. 4, Issue 4, pp 753–762, December 2013.

[4] Qinglei Zhang and Wenying Feng, "Network Intrusion Detection by Support Vectors and Ant Colony", Proceedings of the 2009 International Workshop on Information Security and Application (IWISA 2009), pp 639-642, 2009.

[5] Ashis Pradhan, "SUPPORT VECTOR MACHINE-A Survey", International Journal of Emerging Technology and Advanced Engineering, Vol. 2, Issue 8, pp 82-85, 2012

[6] Marco Dorigo and Christian Blum, "Ant colony optimization theory: A survey", Theoretical Computer Science 344, pp. 243–278, 2005.

[7] Meng Lingxi and Sun Guang, "An Improved Ant Colony Clustering Method for Network Intrusion Detection", 2013 IEEE Eighth International Conference Networking, Architecture and Storage, pp. 312 - 316, July 2013.

[8] Fernando E. B. Otero, Alex A. Freitas, and Colin G. Johnson, "A New Sequential Covering Strategy for Inducing Classification Rules With Ant Colony Algorithms", IEEE Transactions on Evolutionary Computation, Vol. 17, No. 1, pp. 64 - 76, February 2013.

[9] Feng Chen, Xikui Ma, Yanzhen Zhao, and Jianlong Zou, "Support Vector Machine Approach for Calculating the AC Resistance of Air-Core Reactor", IEEE Transactions On Power Delivery, Vol. 26, No. 4, pp. 2407 - 2415, October 2011.

[10] Nada M. A. Al Salami, "Ant Colony Optimization Algorithm", UbiCC Journal, Volume 4, Number 3, pp. 823-826, August 2009.

[11] Zar Chi Su Su Hlaing and May Aye Khine, "An Ant Colony Optimization Algorithm for Solving Traveling Salesman Problem", International Conference on Information Communication and Management IPCSIT vol.16 , pp. 54-59, Singapore-2011.

[12] Kambiz Forqandoost Haqiqi and Tohid Kazemi, "Ant colony optimization approach to portfolio optimization", International Journal of Trade, Economics and Finance, Vol. 3, No. 2, pp. 148-153, April 2012.

[13] Ajith Abraham, Nadia Nedjah and Luiza de Macedo Mourelle, "Evolutionary Computation: from Genetic Algorithms to Genetic Programming", Studies in Computational Intelligence, Studies in Computational Intelligence (SCI) 13, pp. 1–20, 2006.

[14] Mohd Afizi Mohd Shukran Omar Zakaria, Noorhaniza Wahid Ahmad, Mujahid Ahmad Zaidi, "A Classification Method For Data Mining Using SVM-Weight And Euclidean Distance", Australian Journal of Basic and Applied Sciences, Vol. 5, Issue 9, pp. 2053-2059, 2011.

[15] Mrs. A. R. Patil, Dr. S. S. Subbaraman, "A Review On Vision Based Hand Gesture Recognition Approach Using Support Vector Machines", IOSR Journal of Electronics and Communication Engineering (IOSR-JECE), pp. 07-12, 2011.

