Intrusion detection through combining C 4.5 with ant colony networks

Meghana Solanki\textsuperscript{1} & Avinash Godase\textsuperscript{2}
\textsuperscript{1}Assistant Professor, DYPCOE, Pune
\textsuperscript{2}Assistant Professor, GPS, Satara

Abstract: Intrusion Detection system quietly acts as a supervisor for attacks as well as gives notification services and Intrusion Prevention system intensely freezes the threat. In this paper, we recommend a new machine-learning classification algorithm which is applied to network intrusion detection. The primary aim is to classify network activities into to two categories either as a normal or abnormal which depreciate misclassification. There are different classification models which are developed for network intrusion detection. Every model has its strengths as well as weaknesses. Our new approach combines the C 4.5 classification method with CSOACNs to take the profit of both feature’s while bypassing their vulnerability. Our algorithm is implemented as well as assessed using a standard benchmark KDD99 data set. Experiments show that CCAC (Combining C 4.5 with Ant Colony) outperforms C 4.5 alone or CSOACN alone in terms of both classification rate as well run-time efficiency.

1. Introduction

Data clustering and classification have become increasingly significant as well as a challenging area in case of today’s information system management. There are various tools as well as methods have been developed for intrusion detection. Due to the exponential growing-in-size as well as high dimensional data inputs not too many are sufficient and efficient enough for real applications. Computer systems are defended by Intrusion Detection Systems (IDSs) from various cyber attacks and computer viruses. There are two primary presumption in the research of intrusion detection. First is user as well as program activities are examinable by computer systems and second is normal as well as intrusion activities must have explicit behaviors [1]. Network Intrusion Detection Systems (NIDS) acts as a standard component in security infrastructures. Network administrators can detect policy violations by using Network Intrusion Detection Systems. Now a day’s principal focus of research is Anomaly-based network intrusion detection systems. It also gives focus on development in the field of intrusion detection. The behavior of a system is monitored by Anomaly detection systems (ADS). It indicates significant deviations from the normal activity as anomalies. Recently, anomaly detection has been used for identifying attacks in computer networks are identified by anomaly detection. It also identifies malicious activities in computer systems as well as misuses in case of Web systems. There are many recent classes of ADS developed using machine learning techniques such as artificial neural-networks, fuzzy classifiers, multivariate analysis, and others. They have become popular due to their very high detection accuracies at low false positive rate. Artificial neural-networks, pattern matching build anomaly detection methods with single machine learning techniques. Multiple machine learning methods have a better performance yield over individual methods.

2. Literature Survey

Intrusion detection has been studied as a classification problem for long time by making use of machine learning techniques. In this paper [2] author gives review of different approaches. It includes K-Nearest Neighbor (K-NN), Support Vector Machines (SVMs), Decision Trees (DTs), Bayesian, Self-Organized Maps (SOMs), Artificial Neural Networks (ANNs), Generic Algorithms (GAs), and Fuzzy Logic, as well as hybrid classifiers. All these methods improve the performance of the classifier. The review indicates that SVM as well as K-NN were the most widely used techniques. The use of a hybrid increased rapidly after 2004. It also became mainstream. In this paper [3] author has given a thorough survey of intrusion detection by making use of computational intelligence. It has given the details of the classification algorithms as well as swarm intelligence methods. It solves problems by making use of the decentralized agents. In this paper [4] author has combined not only On Line Analytical Processing (OLAP) tools but also data mining techniques. It is proved that the association of the two fields gives a good solution which deal with vulnerabilities of IDSs. It includes low detection accuracy as well as high false alarm rate
vulnerabilities. There are different number of methods were designed using ACO for intrusion detection. In this paper [5] author proposed an ant classifier. Author used more than one colony of ants to detect solutions in case of multiclass classification problem. In this paper [6] author proposed another ant-based clustering algorithm which detect intrusions in a network. They demonstrated that the performance was comparable to some already known classification methods like SVM, DT, and GA. In this paper [7, 8] authors assesses the basic ant-based clustering algorithms. He proposed several improvement blueprints which is useful to overcome the limitations of these clustering algorithms. These algorithms would not function good on clustering large as well as high-dimensional network data. In this paper [9] author applied Support Vector Machine to anomaly detection in the one class setting. Such techniques use one class learning techniques for SVM [10]. Support Vector Machine (SVM) is applicable for supervised intrusion detection in [11]. Complex regions can be learned by Kernels, such as Radial Basis Function (RBF) kernel. In this paper [12] author uses Robust Support Vector Machine (RSVM). It is sturdy to the presence of anomalies in the training data. RSVM have been applied to system call intrusion detection [13].

3. Anomaly Detection with C4.5 Decision Trees

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. Algorithm 1

Given a set X of cases, C4.5 first widens a primary tree using the divide-and-conquer strategy as follows:
1. If all the cases in X belong to the same class or X is small, the tree is a leaf stamped with the most frequent class in X.
2. Otherwise, elect a test based on a single attribute with two or more outcomes. Make this test the root of the tree with one branch for each outcome of the test, segregate X into corresponding subsets X1, X2,... according to the outcome for each case, and apply the same formula recurrently to each subset.

There are usually many tests that could be chosen in this last step. C4.5 uses two heuristic criteria to rank possible tests: information gain, which reduces the total entropy of the subsets {X_i} (but is heavily biased towards tests with numerous outcomes), and the default gain ratio that partitions information gain by the information provided by the test outcomes.

4. Clustering based on a self-organizing ant colony network (CSOACN)

A clustering model based on self-organizing ant colony networks (CSOACN) is methodically scheduled for the purpose of intrusion detection system. We used a nonlinear probability conversion function rather than using the linear segmentation function of the CSI model. It can help to solve linearly inseparable problems. As an ant colony network acquires number of features e.g. flexibility as well as robustness. It also own properties not only decentralization but also self-organization, it can recommend very crucial heuristics. Optimization and control algorithms based on swarm intelligence, including Ant Colony Optimization and Ant Colony Routing, are notable [14]. In the data mining area, there are many researches using the metaphor of ant colonies for purpose of clustering [15]. In order to design a self-organized ant colony network, the problem should be illustrated in term of some standard mathematical objects. In the real world of a self-organized ant colony network, a population of ant-like agents moves substances on a two-dimensional frame so as to clump similar substances into the same zone. Substance and ant are the two primary entries in the problem. When this process going on at that time it is necessary to recognize that whether a substance is identical to its block substances. Swarm sameness is used to explain this property of each substance.

5. Combining the C 4.5 method with CSOACNs

We are now going to introduce the new machine learning algorithm, named Combining C 4.5 with Ant Colony (CCAC). It is based on a combination of the two algorithms discussed above (i.e. C 4.5 and CSOACN). In this algorithm, C 4.5 and CSOACN are two correlative phases which are taken multiple times. C4.5 is used to generate decision tree and that tree is used for classification purpose while a CSOACN is used to find data added to C 4.5 clusters and to finally generate models for normal data as well as for each class of abnormal data. Algorithm for training in CCAC is given below.

Algorithm 2: Training in CCAC

Input: A training data set.
Input: N – number of training iterations.
Input: TT – detection rate threshold.
Output: C 4.5 and CSOACN Classifiers.

1 begin
2 Normalize the data;
3 Let d be the detection rate, initially 0;
4 while rd< TT do
5 for k = 1, · · · , N do
6 C 4.5 clustering phase;
7 Ant clustering phase;
8 end
9 end
Algorithm for IDS using CCAC is given below.

Algorithm 3: Intrusion detection using CCAC
Input: C 4.5 and CSOACN Classifiers from Algorithm 2.
Input: A new data item x.
Output: L – the classification label of x.
1 begin
2 LS ← classification of x with C 4.5;
3 LC ← classification of x with CSOACN;
4 if LS = LC = normal then
5 L ← normal;
6 else if LS <> LC then
7 L ← amphibolous;
8 else
9 L ← LC ;
10 (a sub-class of abnormal is detected by the CSOACN abnormal classifier.)
11 end

6. System design
The comprehensive architecture of the new IDS is shown in Figure 1. The architecture made up of five functional elements which are shown in oval shape in the figure. Its several storage elements are shown in rounded-corner rectangle shape in the figure. The data sets are delineated in rectangle. C 4.5 and CSOACN use different forms of input data. The training data set is transposed dynamically in both algorithms. There are three types of storage that are distributed in the new IDS. One of this denoted as Disk which is used for storing all the candidate data for training. Second one is denoted as cache1 which is used for storing the data of each C 4.5. Third one is denoted as cache2 which is used for storing the training data of each CSOACN period. The types of data stored in these three storages are represented as point, Lagrange point, and object, respectively. We have selected the KDD99 data set. KDD data set is used for training and testing of the newly created IDS. In KDD99, attacks are divided into four main categories first is Denial of Service (DoS), second is Remote to Local (R2L), third is User to Root (U2R), and 4th is Probing. In order to differentiate normal connections from attacks, 41 higher-level features are defined in network connection data item in KDD99.

7. Comparisons and discussions
The proposed CCAC algorithm outperforms pure C 4.5. It gives with higher average detection rate as well as less training time. It also gives lower rates of not only false negative but also false positive. It is better than pure CSOACN in terms of less training time. It is also better in term of false alarm rates. Our experiments also demonstrate that the effectiveness of newly created algorithm is commensurable to that of the KDD99 winner. Therefore, CCAC can be considered as an excellent method that outperforms its counterpart methods. The proposed CCAC algorithm has a very less overall training time. The reason for this is that it gives usage of the data points encircling the boundaries of two classes for the construction of the classifier. The new intrusion detection system based on CCAC can be applied in different cases by constructing the detection classifier. C 4.5 is reasonable for the time intense case. On the other hand, the CSOACN training mode is good enough for the incisiveness intensive case. It can solve multiclass problems upon both labeled as well as unlabeled data. The performance of IDS can be balanced by the CCAC mode in terms of not only efficiency but also accuracy.
8. Conclusions

In this paper, we have recommended a new algorithm (CCAC) for developing classifiers with clustering. It is applied to the intrusion detection problem. This approach integrates two existing machine learning methods i.e. C 4.5 as well as C4OACN. By using combined approach we can achieve better performance not only in detection accuracy rate but also in faster running time.

9. References


