A New Intrusion Detection Method Based on Data-Oriented Classification of Attacks

TAO ZOU[1], HUA CHEN [1],CUI ZHANG[2], MINHUAN HUANG [1]
Beijing Institute of System Engineering, P.O.Box 9702-19, Beijing, PRC China 100101
Beijing Graphics Institute, Beijing, China, 100029

Abstract: - The most acute problem for misuse detection method is its inability to detect new kinds of attacks. A new detection method based on data-oriented classification of attacks is proposed to solve this problem. After analyzing its significance, a practical scheme which uses relevant feature subset codes clustering is designed. Applying Concept Hierarchy Generation for attack Labels (CHGL), inductive learning algorithms can learn attack profiles on high concept levels. Experimental results show the advantage of this method.

Key-words: - intrusion detection systems, misuse detection, classification, clustering.

1. Introduction

Intrusion Detection (ID) approaches can be categorized into anomaly detection and misuse detection. Misuse detection methods attempt to model attacks as specific patterns, and then systematically scan the system for occurrences of these patterns. It can detect known attacks fairly reliably with a low false positive rate. However, novel attacks or even variants of common attacks often go undetected [1]. This is because its detection engine cannot create the profiles of new attack types at detection stage. Artificial neural network, data mining and rule inductive learning, etc have been employed to automatically model an Intrusion Detection System (IDS) [2-4]. We also focus on the automatically modeling technology. The current research is to make IDS detect more instances, including new kinds of attacks.

A new idea of learning target concepts on variant concept levels based on traditional misuse detection method is proposed in this paper. First, a clustering algorithm is applied to relevant feature subset codes to output a detection-oriented classification of attacks. Then Concept Hierarchy Generation for Labels (CHGL) is executed. Finally, re-organized training data is used to train common machine learning algorithm to obtain high concept level attack profiles. This method, called Concept Level Misuse Detection (CLMD), has the ability to detect new kinds of attacks.

2. Methodology

2.1 Learning target concept on variant levels

Machine learning involves inductive learning, which acquires general concepts from a large amount of specific training examples and then uses these concepts to predict new instances. A concept hierarchy is a sequence of mappings from a set of low-level concepts to higher level and more general ones [5].

Target concepts can be defined on variant levels in a concept hierarchy. For example, biology can be defined as a standardized 7-level hierarchy, from Species up to Kingdom, arranged in a hierarchical order [6]. The concepts learned from the same example set are completely different if inductive learning is executed on different levels. Animalia is a concept on the top level while catenifer is on the lowest. The amounts of instances covered by these concepts are also much different. The higher the level is, the more instances it can cover. Furthermore, if a target concept is learned on high-level, it is possible for the instances which belong to unknown subcategories to be correctly predicted into their high-level categories. This is just what traditional misuse detection methods need.

2.2 The principle of classification for ID

A concept hierarchy can be constructed from classification. Classification groups instances into higher level taxa based primarily on apparent resemblance or on the possession of shared traits. Instances can be viewed in many different ways. For example, IDS can be classified as host based IDS and network based IDS according to its configuration and data source. But in respect of detection method, IDS

can also be categorized into anomaly detection system and misuse detection system.

No matter on what respect a classification is based, it needs to obey some basic criteria to ensure that the resulting classification is sound. However, the choice of these basic criteria is not unambiguous. Mutually exclusive, exhaustiveness, etc, are common criteria for most classifications. But when we take into account machine learning and ID application, the most important criterion is neither of these two but the strong resemblance between instances of the same class. When a concept is being learned, the stronger this resemblance is, the simpler the hypothesis space *H* will be. That means the size of H is small and we need a classifier with only small capacity.

The followings are some of the conclusions from Statistical Learning Theory (SLT) [7]:

$$VC(H) \le \log_2 |H| \quad (1)$$

$$R(D_n) \le R_{emp}(D_n) + \sqrt{\frac{VC(H)(\ln(2n/VC(H)) + 1) - \ln(\eta/4)}{n}} \quad (2)$$

where $|H|$ is the size of hypothesis space H and *VC(H)* means VC dimension. *VC* dimension is a measure of the capacity of a learning algorithm, and it is the core concept in SLT. $R(D_n)$ and $R_{emp}(D_n)$ denote expected risk and empirical risk respectively. The right side of Eq. (2) is called "risk bound" and its second term is called "VC confidence". It is obvious that reducing the size of *H* can improve the generalization capability of a classifier.

But when applying machine learning technology in the practice of ID, we will find that this important criterion is not easy to meet. Instances in training dataset are often expressed as <*X, Y*>, where *X* is feature vector and *Y* is class label. Though there may exist many classification methods, nearly no one is consistent with the criterion of strong resemblance within class. This is because these classifications are based on variant criteria but not focus on data resemblance. Actually, the resemblance between instances of the same class, with respect to *X*, is the key point of improving generalization capability.

2.3 Related research

Attack classification is the foundation of ID. Through the years, many classifications of attacks have been presented, some concentrated on the intruders and their methods while others on the characteristics of the computer system which make the intrusion possible. Early in 1974, Lackey [8] presented six categories of penetration techniques based on many examples of actual system penetration. Neumann and Parker categorized computer misuse techniques into nine classes [9]. Brinkley and Schell [10] categorized what they call information-oriented computer misuse (regarding the security aspects confidentiality and integrity, but not availability, which the authors called resource-oriented computer misuse) into six different classes.

In 1995, Kumar made a classification of intrusions according to attack signatures they left in the audit trail of the system [11]. This classification provided a different way of viewing ID, namely in terms of the types of patterns that can be used to detect intrusions, instead of the generic anomaly and misuse approaches. Though this classification can be viewed as a detection-oriented taxonomy, it has the following problems: 1) It is lack of concept hierarchy and we cannot learn high-level concepts from it. 2) It is not data-oriented. Different IDS uses different set of features and different analytical models. Only data-oriented classification is helpful to detect intrusions. Beside, the author himself also pointed out that signature analysis assumed the integrity of event data. Thus, attacks involving spoofing which produce the same events cannot be reliably detected. 3) All attacks must be classified by hand. That is laborious and inefficient. All of the classifications mentioned above cannot meet the criterion we analyzed in Section 2.2. In order to use machine learning on high concept level to improve generalization capability, data-oriented classification needs to be studied.

2.4 Concept level misuse detection

Raising the level of target concepts can help the system to predict more instances of attacks, including those belonging to new attack types. In this paper, a new ID method based on data-oriented classification of attacks, named Concept Level Misuse Detection (CLMD), is proposed. CLMD uses the same training dataset as misuse detection system, in which the class labels of the examples, such as Land and Smurf, are always on low level.

Instances are represented in the form of <*X, Y*>. For a special kind of attacks, only some of the features are relevant to it. Different Relevant Feature Subsets (RFS) are needed to detect different kinds of attacks. So, a Feature Subset Selection (FSS) algorithm is used first to get the Relevant Features (RF) for each kind of attacks and output a RF dataset. Each individual in this dataset is in the form of <*RFS, Attack type*>. In order to make sure that all the features in RFS are the most relevant to its corresponding attack type, a

Wrapper method is used instead of a Filter one [12].

To a certain extent, RFS denotes the data characteristic of attacks. If some two RFSs are closer in feature space than others, the two corresponding attack types will have more similarities than others. So, attack classification according to data resemblance can be made by RFS clustering. If the total number of features for each instance is m, RFS can be encoded into an m dimension code vector whose element is 0 or 1.

$$co_i = \begin{cases} 1, & \text{iff } f_i \text{ is } RF \\ 0, & \text{else} \end{cases} \quad (3)$$

$F = (f_1, f_2, \cdots, f_m)$ is the full feature set and $(co_1, co_2, \cdots, co_m)$ is the RFS code. Distance between point a^1 and a^2 can be defined as:

$$d(a^1, a^2) = \left(\sum_{i=1}^{m} \left| co_{a^1 i} - co_{a^2 i} \right|\right) / \left(\sum_{i=1}^{m} \left| co_{a^1 i} + co_{a^2 i} \right|\right) \quad (4)$$

where $co_{a^1 i}$ is the value of the ith scalar in the RFS code for attack type a^1.

If there are totally n kinds of attacks in training dataset, there will be k clusters after clustering, and $k \le n$. A hierarchical cluster tree is created using linkage algorithm. A threshold which determines how the cluster function creates clusters can be set in two ways: the threshold for the inconsistency coefficient or the maximum number of clusters to retain in the hierarchical tree.

After the clusters are produced, all labels of attacks are replaced by the cluster-labels and the training data is reorganized as the input of an induction algorithm. Because the cluster-labels denote the target concepts on high-level, this process is named as Concept Hierarchy Generation for Labels (CHGL). The level of CHGL can be agilely controlled by setting different threshold value in cluster algorithm. CHGL focuses on the data resemblance, so this classification is data-oriented.

Clustering instances according to their feature-vector values directly can produce data-oriented classification too. But compared with CHGL, this method has some difficulties. It will be very complex when there are a large number of instances in training dataset. Furthermore, irrelevant features will greatly influence system performance, and the value of weight for each feature is difficult to determine. Liang uses this method to detect intrusions and the detection rate is less than 70% [13].

3. System evaluation and experimental results

The dataset we use comes from 1999 KDD intrusion detection contest. It is a version of dataset in 1998 DARPA Intrusion Detection Evaluation Program [14]. The raw training data was about four gigabytes of compressed binary TCP dump data from seven weeks of network traffic and was processed into about five million connection records. Similarly, the two weeks of test data yielded around two million connection records.

There are 22 and 37 attack types in training dataset and testing dataset respectively. Every instance has 41 features. Average linkage uses the average distance between all pairs of objects in group i and group j as the measurement of proximity between these two groups. When threshold is set to 0.8, 15 clusters are outputted. Among them, cluster 14 and 15 are the largest, each contains three attack types. Cluster 11 to 13 each contains two attack types. The remaining ten attack types form 10 clusters because there is no similar type for each of them. The result is shown in Fig. 1.

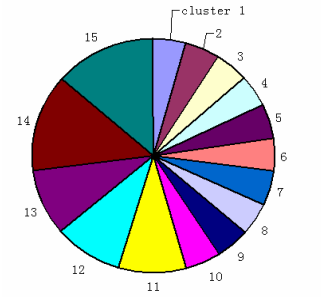

Fig.1. The result of clustering

After CHGL, the original training dataset is converted into a new one which contains only 15 clusters. A rule induction algorithm named RIPPER [15] is used to get the attack profiles to predict new instances.

Table 1 shows the performance of CLMD compared with traditional misuse detection system. With the cost of only 4 false alarms increasing, CLMD can detect 1433 more attacks and two kinds of new attack types, Httptunnel and Saint, in testing dataset. The ratio of performance improvement to degradation is about 100 with respect to detection rate and false positive rate. The size of hypothesis also reduces about 11%. This simplification will be helpful to speedup the matching process at detection stage.

	Detected new attack types	Detected	Detection rate	Number of false positive	False positive rate	Hypothesis (rules / conditions)
Misuse	—	225472	90.03%	301	0.497%	98 / 431
CLMD	Httptunnel, Saint	226905	90.60%	305	0.503%	85 / 386
Differences	2 kinds↑	1433↑	0.57%↑	4↓	0.006% ↓	13 / 45↑

↓: Performance degradation ↑: Performance improvement

Table 1. System performance of CLMD and misuse detection

	Instances	Detected	Detection rate
Saint	736	627	85.19%
Httptunnel	158	106	67.09%

Table 2. New attack types detected

Table 2 shows the detection rate of Httptunnel and Saint. Both detection rates are higher than 67%. This is fairly high for unknown attack types.

4. Discussion and Conclusion

This paper puts forward a new ID method called CLMD. It can detect more attack instances including those belonging to new attack types with the help of a data-oriented classification, which outputs a concept hierarchy. Experimental results have shown the improvement of the system performance. Another advantage of this method is that attack types are automatically classified by computer, not by human.

One thing we must emphasize is that the dataset we use in the experiment is not collected for the purpose of testing CHGL. The number of attack types in training dataset is not adequate enough, which is even fewer than that of features. That means the attack types in training dataset are sparsely distributed and not all the resemblances between the instances of the same cluster are very strong. The results shown above are just primary. In practice, there are plentiful attack types and the advantage of this method will be more distinct.

References:

[1] A. Ghosh, J. Wanken, and F, Charron. Detecting Anomalous and Unknown Intrusions Against Programs. In *Proceedings of the 14th Annual Computer Security Applications Conference* December 7-11, 1998 Phoenix, Arizona.

[2] J. Ryan, M. Lin, and R. Miikkulainen, Intrusion Detection with Neural Networks, in *the AAAI Workshop*, 1997, pp. 72-79.

[3] W. Lee, S. J. Stolfo, and K. W. Mok, A Data Mining Framework for Building Intrusion Detection Models, in *Proceedings of the 1999 IEEE Symposium on Security and Privacy*, 1999, pp. 120-132.

[4] Guy Helmer, Automated Discovery of Concise Predictive Rules for Intrusion Detection. *Journal of Systems and Software*, Vol.60, February 2002.

[5] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, New York, 2000.

[6] Donald L. Blanchard, The ABC's of Animal Taxonomy. *The Cold Blooded News,* Vol.26, No.1, January 1999.

[7] Vapnik V., *The Nature of Statistical Learning Theory (the second edition)*. New York: Springer-Verlag 1998.

[8] R.D. Lackey, Penetration of Computer Systems, an Overview. *Honeywell Computer Journal,* 8(2): 81–85, 1974.

[9] P.G. Neumann and D.B. Parker, A Summary of Computer Misuse Techniques. In *Proceedings of the 12th National Computer Security Conference,* pages 396–407, Baltimore, Maryland, USA, Oct. 10–13, 1989.

[10] D.L. Brinkley and R.R. Schell, What Is There to Worry About? An Introduction to The Computer Security Problem. In M.D. Abrams, S.Jajodia, and H.J. Podell, editors, Information Security: An Integrated Collection of Essays, pages11–39. IEEE Computer Society Press, 1995.

[11] S.Kumar, Classification and Detection of Computer Intrusions. PhD thesis, Purdue University, West Lafayette, Indiana, USA, Aug. 1995.

[12] Tao Zou, and Hongwei Sun, Data Reduction in Network Based Intrusion Detection System, *Journal of National University of Defense Technology*, 2003.

[13] Liang Tie-zhu, Li Jian-Cheng, Wang Ye, A Novel Clustering-Based Method to Network Intrusion Detection. *Journal of National University of Defense Technology*, Vol.24, No.2 2002.

[14] DARPA 1998 Intrusion Detection Evaluation, in http://www.ll.mit.edu/IST/ideval/index.html.

[15] Cohen, W.W., Fast Effective Rule Induction. In: Proceedings of *the 12th International Conference on Machine Learning*, Lake Tahoe, CA. Morgan Kaufmann, Los Altos, 1995.