A Comprehensive Study in Data Mining Frameworks for Intrusion Detection

R.Venkatesan[1], R. Ganesan[2], A. Arul Lawrence Selvakumar[3]
Research Scholar/CS, CMJ University, Shillong–793003, Meghalaya[1]
Professor/ E& I Engineering, N I C E, Kumaracoil, Tamilnadu[2]
Professor & Head, Department of CSE, Rajiv Gandhi Institute of Technology, Bangalore, Karnataka[3]

Abstract

Intrusions are the activities that violate the security policy of system. Intrusion Detection is the process used to identify intrusions. Network security is to be considered as a major issue in recent years, since the computer network keeps on expanding every day. An Intrusion Detection System (IDS) is a system for detecting intrusions and reporting to the authority or to the network administration. Data mining techniques have been successfully applied in many fields like Network Management, Education, Science, Business, Manufacturing, Process control, and Fraud Detection. Data Mining for IDS is the technique which can be used mainly to identify unknown attacks and to raise alarms when security violations are detected. The purpose of this survey paper is to describe the methods/ techniques which are being used for Intrusion Detection based on Data mining concepts and the designed frame works for the same. We are also going to review the related works for intrusion detection.

Keywords

Data Mining, Intrusion Detection System (IDS), Network Security, Misuse Detection, Anomaly Detection, Classification, Clustering, MADAM ID, ADAM, JAM

1. Introduction

In recent years, many researchers are focusing to use Data Mining concepts for Intrusion Detection. Data mining is a process to extract the implicit information and knowledge. In the other hand intrusions are the activities that violate the security policy of the system, and intrusion detection is the process used to identify intrusions. In this paper, we are going to study briefly about how data mining concepts are used to develop various frameworks/models like ADAM, JAM and MADAM ID.

1.1 Data Mining

Data mining is the process of discovering interesting patterns (or knowledge) from large amounts of data. The data sources can include databases, data warehouses, the Web, any other information repositories, or data that are streamed into the system (dynamically). The Knowledge Discovery in Databases (KDD) Process is used to denote the process of extracting useful knowledge from large data sets. The KDD process involves a number of steps and is often interactive, iterative and user-driven decision making rules [1]. Data mining is the most vital step in the KDD process, and it applies *data mining techniques* to extract patterns from the data.

- *Know the application domain*: to understand the back ground of the knowledge and to specify the goal.
- *Data Collection*: includes creating a target dataset which is relevant to the analysis
- *Data Mining:* applying an appropriate algorithm to extract useful information using techniques.
- *Data Interpretation:* to understand the discovered patterns and to confirm the goal is achieved.
- *Knowledge Representation*: the final stage of representing the discovered knowledge.

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks and it can be classified into two categories:

- **Descriptive:** to characterize the general properties of data in the database
- **Predictive:** to perform inference on data and to make predictions

1.2 Intrusion Detection

The Intrusion Detection concept was introduced by Anderson J. in 1980 [2]. Intrusion detection systems (IDSs) are usually deployed along with other preventive security mechanisms, such as access control and authentication, as an argument that protects information systems. IDSs may be classified into:

i. *Host-based IDSs* - Host based IDSs examine data held on individual computer that serve as hosts. The network structural design of host based is an agent-based, which means that software resides on each of the hosts that will be governing by the system.

ii. *Distributed IDSs* –it gathers audit data from multiple hosts and possibly the network that connects the hosts, aiming at detecting attacks involving multiple hosts.

iii. *Network-Based IDS* - Uses network traffic as the audit data source, relieving the burden on the hosts that usually provide normal computing services and detect attacks from network.

Denning D.E in 1986 [3] presented the first intrusion detection model, which has six main components:

- *Subjects* refer to the initiators of activity in an information system; they are usually normal users.
- *Objects* are the resources managed by the information system, such as files, commands and devices.
- *Audit records* are those generated by the information system in response to actions performed or attempted by subjects on objects. Examples include user login, command execution, etc.
- *Profiles* are structures that characterize the behavior of subjects with respect to objectives in terms of statistical metrics and models of observed activity.
- *Anomaly records* are indications of abnormal behaviors when they are detected.
- *Activity rules* specify actions to take when some conditions are satisfied, which updates the profiles, detects abnormal behaviors, relate anomalies to suspected intrusions, and produce reports.

2. Intrusion Detection Techniques

The intrusion detection techniques based on data mining [4], [5] are generally falls into one of two categories: *anomaly detection* and *misuse detection*.

2.1 Anomaly Detection

Anomaly detection attempts to determine whether deviation from an established normal behavior profile can be flagged as an intrusion [6].Anomaly detection consists of first establishing the normal behavior profiles for users, programs, or other resources of interest in a system, and observing the actual activities as reported in the audit data to ultimately detect any significant deviations from these profiles. Most anomaly detection approaches are statistical in nature. Anomaly detection may be divided into *static* and *dynamic anomaly detection*. A static anomaly detector is based on the assumption that there is a portion of the system being monitored that does not change. The static portion of a system is the code for the system and the constant portion of data depends upon the correct functioning of the system. Dynamic anomaly detection typically operates on audit records or on monitored networked traffic data. An audit record of operating systems does not record all events; they only record events of interest. Strength of anomaly detection is its ability to detect previously unknown attacks.

2.2 Misuse Detection

Misuse detection works by searching for the traces or patterns of well-known attacks. Lee et al. [7] designed a signature-based database intrusion detection system (DIDS) which detects intrusions by matching new SQL statements against a known set of transaction fingerprints. Misuse detection is considered complementary to anomaly detection. This system usually searches for patterns or user behavior that matches known intrusion or scenarios, which are stored as signatures. If a pattern match is found, it signals an event then an alarm is raised. Pattern, Data mining, and state transition analysis are some of the approaches of misuse detection. To perform this detection method, each scenario need to described or modeled. Misuse detection is based on extensive knowledge of patterns associated with known attacks provided by human experts.

2.3 Pros and Cons of Anomaly and Misuse Detection

Table 1: Pros and Cons of Anomaly Detection and Misuse Detection

Technique	Pros	Cons
Anomaly Detection	Is able to detect unknown attacks based on audits	High false-alarm and limited by training data.
Misuse Detection	Accurately and generate much fewer false alarm	Cannot detect novel or unknown attacks

2.4 Combining Misuse and Anomaly Detection

Anomaly detection and misuse detection have major shortcomings that hamper their effectiveness in detecting intrusions. Research can be carried into intrusion a detection methodology which combines both the anomaly detection approach and the misuse detection approach [8]. The combined approach permits a single intrusion detection system to monitor for indications of external and internal attacks. Pattern recognition possesses a distinct advantage over anomaly and misuse detection methods in that it is capable of identifying attacks which may occur over an extended period of time, a series of user sessions, or by multiple attackers working in concert.

2.5 Drawbacks of current IDS

Intrusion Detection Systems (IDS) has become a standard component in security infrastructures as they allow network administrators to detect any violations. These security violations range from external attackers trying to gain unauthorized access to insiders abusing their access. Current IDS have a number of significant drawbacks [9]:

- **False Positives** – A common complaint is the amount of false an IDS will generate.
- **False Negatives** – In this case, IDS does not create a signature or alarm, when an intrusion is actually happened.
- **Data Overload** – In this aspect, Misuse Detection cannot be related directly, however, it is very important to analyze how much data an analyst can efficiently and effectively analyze.

2.6 Need of using data mining approaches in IDS

A team in Minnesota University (1990) recognized the need for existence of standardized dataset to train IDS tool. Minnesota Intrusion Detection System (MINDS) combines signature based tool with data mining techniques. Signature based tool (Snort - freeware) are used for misuse detection & data mining for anomaly detection. The reasons for using Data Mining approaches in IDS are:

1. It is very difficult to build IDS using programming languages, which requires more explicit data and functional knowledge.

2. The reliability, compatibility and dynamic nature of machine-learning make it a suitable solution for this situation.

3. The environment of an IDS and its classification task highly depend on user-driven preferences.

3. Data Mining approaches

Data mining generally refers to the process of (automatically) extracting models from large stores of data [10]. The recent rapid development in data mining has made available a wide variety of algorithms, drawn from the fields of statistics, pattern recognition, machine learning, and database. There are several types of algorithms [10] which are particularly related to intrusion detection.

- **Classification:** classifies a data item into one of several pre-defined categories. These algorithms normally output "classifiers". An ideal application in intrusion detection would be to gather sufficient "normal" and "abnormal" audit data for a user or a program, then apply a Classification algorithm to learn a classifier that can label or predict new unseen audit data as belonging to the normal class or the abnormal class.
- **Link analysis:** determines relations between fields in the data base records. Correlations of system features in audit data, for example, the correlation between command and argument in the shell command history data of a user, can serve as the basis for constructing normal usage profiles.
- **Sequence analysis**: models sequential patterns. These algorithms can discover what time-based sequences of audit events are frequently occurring together. These frequent event patterns provide guidelines for incorporating temporal and statistical measures into intrusion detection models.

3.1 Association Rule or Dependency Mining

Association analysis is the discovery of association rules showing attribute – value conditions that occur frequently together in a given set of data. Association analysis widely used in transaction data analysis. This approach work on data dependency, in which one item is modify another item refer with this also modify. The concept of Association mining is to find all co-occurrences relationship called associations. Association Mining has been used in various domains and many efficient algorithms, extensions and applications have reported. In general, Association analysis has been considered as an unsupervised technique, so it can be applied for KDD task. *Agrawal et al., 1993* [11] stated the association rule.

Algorithm: let A be a set of attributes, and I be a set of values on A, called items. Any subset of I is called an item set. The number of *items* in an item set is called its length. Let D be a database with *n* attributes (columns). Define support (X) as the percentage of transactions (records) in D that contain item set X. An association rule is the expression

$$X \rightarrow Y, [\textit{confidence}, \textit{support}]$$

Here X and Y are item sets, and X ∩ Y = θ. *support*(X U Y) is the support of the rule, and $\frac{support(X \cup Y)}{support(X)}$ is the confidence of the rule.

Apriori Association Rules Algorithm [Agrawal and Srikant, 1994]. Apriori is an algorithm for mining frequent item sets for Boolean association rules [12]. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent item set properties.

Apriori algorithm employs an iterative approach known as a level wise search, where k-item sets are used to explore (k + l)-item sets.

```
scan database D to form L1 = {frequent 1-itemsets};
k =2; /* k is the length of the item sets */
While Lk−1 ≠ θ do begin /* association generation */
    for each pair of lk-11, lk-12 Є Lk−1 and lk-11 = lk-12
    where
        their first k −2 items are the same do begin
            construct candidate item set ck such that its first k −2 items are the same as lk-11, and the last two items are the last item of lk-11 and the last item of lk-12;
            if there is a length k −1 subset sk−1 Є ck and sk−1 ∉ /Lk−1 then remove ck; /* the prune step */
            else
                add ck to Ck;
        end
    scan D and count the support of each ck Є Ck;
    Lk = {ck|support(ck) ≥ minimum_support};
    k = k +1;
end
for all lk,k > 2 do begin /* rule generation */
    for all subset am Є lk do begin
        conf = support(lk)/support(am);
        if conf ≥ minimum_confidence then begin
            output rule am →(lk − am),
            with confidence = conf and support = support(lk);
        end
    end
end
```

3.2 Classification

Classification is the method of identifying a set of categories (sub-populations) with a new observation belonging, on the basis of training set of data containing observations (or instances) whose category membership is known. An algorithm that implements classification, especially in a concrete implementation, is known as a *classifier* [13]. The term "classifier" sometimes also refers to the mathematical function, implemented by a classification algorithm that maps input data to a category. Classification can be thought of as two separate problems – *Binary* and *Multiclass* classification in binary classification, a better understood task, only two classes are involved, whereas multiclass classification involves assigning an object to one of several classes [14]. This is a supervised learning. The class will be predetermined in training phase.

3.3 Clustering- maps data items into groups according to similarity or distance between them. There are many clustering methods available, and each of them may give a different grouping of a dataset. The choice of a particular method will depend on the type of output desired. In general, clustering methods may be divided into two categories based on the cluster structure which they produce [15].

3.3.1 Non Hierarchical - This method divides a dataset of N objects into M clusters, with or without overlap. These methods are sometime divided into partitioning methods, in which the classes are mutually exclusive, and the less common clumping method, in which overlap is allowed. Each object is a member of the cluster with which it is most similar however the threshold of similarity has to be defined.

3.3.2 Hierarchical – Connection Oriented - This method produces a set of nested clusters in which each pair of objects or clusters is progressively nested in a larger cluster until only one cluster remains. The hierarchical methods can be further divided into agglomerative or divisive methods. In agglomerative methods, the hierarchy is build up in a series of *N-1* agglomerations, or Fusion, of pairs of objects, beginning with the un-clustered dataset. The less common divisive methods begin with all objects in a single cluster and at each of N-1 steps divide some clusters into two smaller clusters, until each object resides in its own cluster.

4. Data Mining Frameworks

In the recent years, many research projects have applied data mining techniques for intrusion detection. Here, we survey a representative cross section of these projects. The objective of this survey is to give the overview of the frameworks/models that has been designed to detect interference using data mining.

4.1 ADAM (Audit Data Analysis and Mining)

ADAM proposes applying data mining techniques to discover abnormal patterns in large amounts of audit data [16] [17]. ADAM is flexible to provide the network traffic pattern and uncover some unknown patterns of attacks that cannot be detected by other techniques. ADAM uses several data-mining-related techniques to help detect abnormal network activities.
Type 1: Using Association Rule: Given a set *I* of items, an association rule is a rule of the form *X*→*Y*, where *X* and *Y* are subsets (called item sets) of *I* and X∩Y = φ. Association rules are usually discovered from a set *T* of transactions, where each transaction is a subset of *I*. The rule *X*→*Y* has a *support s* in the transaction set *T* if *s*% of the transactions in *T* contain *X* and *Y*, and it has a *confidence c* if *c*% of the transactions in *T* that contain *X* also contain *Y*. However, ADAM doesn't use association rules directly; instead, it adopts the item sets that have large enough support (called *large item sets*) to represent the pattern of network traffic.
Type 2: Domain-level mining, In this type the system tries to generalize the event attribute values used to describe a network event, after that it discovers large item sets using the generalized attribute values.
Type 3: Classification algorithms can be used to classify large data set and it is quite effective to reduce false alarms. ADAM is an anomaly detection based model and its limitation is it can detect an attack only when it involves a relatively large number of events during a short period of time.

4.2 JAM (Java Agent for Meta-learning)

Stolfo et al., 1997[18] at Columbia University developed JAM, an agent-based distributed data mining framework, and applying it to the problem of credit card fraud detection. The main research effort is on using meta-learning to combine multiple base classifiers separately learned from distributed databases. This framework uses a "supervised" learning approach to build attack models and it requires records of the training datasets used for building attack modes are labeled as either normal or intrusive.

4.3MADAMID (Mining Audit Data for Automated Models for Intrusion Detection)

Lee & Stolfo [19] proposed the advanced state-of-the-art knowledge of intrusion detection by introducing the MADAMID, framework that helps generate intrusion detection models automatically. MADAM ID is used to pre-process raw audit data into records with a set of "intrinsic" features, and then Data mining algorithms are applied to compute the frequent activity patterns from the records. These activities are automatically analyzed to generate an additional set of features for intrusion detection purposes. This type of classification task is commonly referred to as supervised learning as class label for each given training datasets. The Limitations of MADAM ID are:

- It will be applied only at connection level.
- Within connection classification of contents are very challenging

4.4 Related Works in Intrusion Detection

Many researches applied/implemented data mining techniques to design/model IDS. The detailed reports of such developments can be studied in the literatures [3] [20] - [26].

5. Conclusion

Data mining methods are capable of extracting patterns automatically and adaptively from a large amount of data. Various methods related to intrusion detection system are studied briefly. This survey paper states the methods and techniques of data mining to aid the process of Intrusion Detection and the frameworks which were developed using these concepts. The concept of intercepting these two different fields, gives more scope for the research community to work in this area. New approaches will enhance the existing interference detecting system and it will be a stepping stone to build effective and efficient IDS to detect different type of attacks.

References

[1] Fayyad, U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. The KDD process of extracting useful knowledge from Volumes of data. Communications of the ACM, 39(11):27–34, November 1996.

[2] Anderson J., "Computer Security Threat Monitoring and Surveillance," February 26, 1980- revised April 15, 1980.

[3] Dorothy E. Denning. "An Intrusion-Detetcion Model", 1986 IEEE Computer Society Symposium on Research in Security and Privacy, pp 118-31.

[4] Daniel Barbara, Ningning Wu and Sushil Jajodia Detecting novel network intrusion using bayes estimators. In Proceedings of First SIAM Conference on data mining Chicago, 2001.

[5] Bloedorn, Eric, Alan D. Christiansen, William Hill, Clement Skorupka, Lisa M. Talbot, and Jonathan Tivel. Data mining for network

intrusion detection: How to get started. MITRE Technical Report, 2001.

[6] W. Lee and S. J. Stolfo. Data mining approachesfor intrusion detection, In Proceedings of the 7thUSENIX Security Symposium, San Antonio, TX, January 1998.

[7] S.Y. Lee, W. L. Low and P. Y. Wong, “Learning Fingerprints for a Database Intrusion Detection System”, In Proceedings of the 7th European Symposium on Research in Computer Security, Pages 264-280, 2002.

[8] Lunt, T.F. (1989), Real -Time Intrusion Detection. Proceedings from IEEE COMPCON.

[9] SANS: FAQ: Data Mining in Intrusion Detection http://www.sans.org/security-resources/idfaq/data_mining.php.

[10] W. Lee, S.J. Stolfo, K.W. Mok, Algorithms for Mining System Audit Data, in Proc. KDD, 1999.

[11] Agrawal et al 1993] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In Proceedings in the ACM SIGMOD Conference of Management of Data, pages 207-216, 1993.

[12] Agrawal R. and Srikant R., “Fast algorithms for mining association rules,” in Proceeding 20th VLDB Conference, Santiago, Chile, pp. 487–499, 1994.

[13] J.R.Quinlan. C4.5: Programs for machine learning.Morgan Kaufman Publishers, 1993.

[14] Har-Peled, S., Roth, D., Zimak, D. (2003) "Constraint Classification for Multiclass Classification and Ranking." In: Becker, B., Thrun, S., Obermayer, K. (Eds) Advances in Neural Information Processing Systems 15: Proceedings of the 2002 Conference, MIT Press. ISBN 0-262-02550-7.

[15] Manish Joshi / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 2, Mar-Apr 2012, pp.961-964.

[16] Barbara,D., Couto, J., Jajodia, S., & Wu, N. (2001). ADAM: A testbed for exploring the use of data mining in intrusion detection, ACM SIGMOD Record, 30 (4), 15--24.

[17] Barbara, D., Couto, J., Jajodia, S., & Wu, N. (2002). An architecture for anomaly detection, In D. Barbara & S. Jajodia (Eds.), Applications of Data Mining in Computer Security (pp. 63--76). Boston: Kluwer Academic.

[18] S. J. Stolfo, A. L. Prodromidis, S. Tselepis, W. Lee, D. W. Fan, and P. K. Chan. JAM: Java agents for meta-learning over distributed databases. In Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, pages 74–81, Newport Beach, CA, August 1997, AAAI Press.

[19] Lee, W., & Stolfo, S.J. (2000) A framework for constructing features and models for intrusion detection systems. ACM Transactions on Information and System Security, 3 (4) (pp. 227-261).

[20] Teng.H.S, Chen.K and Lu.S.C, “Adaptive Real-Time Anomaly Detection using Inductively Generated Sequential Patterns, in the Proceedings of Symposium on research in Computer Security & Privacy, IEEE Communication Magazine,1990, pp-278-284.

[21] Norouzian.M.R, Merati.S, “Classifying Attacks in a Network Intrusion Detection System Based on Artificial Neural Networks”, in the Proceedings of 13th International Conference on Advanced Communication Technology(ICACT), 2011,ISBN:978-1-4244-8830-8,pp-868-873.

[22] Sekeh.M.A,Bin Maarof.M.A, “Fuzzy Intrusion Detection System Via Data Mining with Sequence of System Calls”, in the Proceedings of International Conference on Information Assurance & security (IAS)2009,IEEE Communication Magazine, pp- 154-158.

[23] Amir Azimi, Alasti, Ahrabi, Ahmad Habibizad Navin,Hadi Bahrbegi, “A New System for Clustering & Classification of Intrusion Detection System Alerts Using SOM”, International Journal of Computer Science & Security, Vol: 4, Issue: 6, pp-589-597, 2011.

[24] Alan Bivens, Chandrika Palagiri, Rasheda Smith, Boleslaw Szymanski, ”Network-Based Intrusion Detection Using Neural Networks”, in Proceedings of the Intelligent Engineering Systems Through Artificial Neural Networks, St.Louis, ANNIE-2002, and Vol: 12, pp- 579-584, ASME Press, New York.

[25] Dewan Md, Farid, Mohammed Zahidur Rahman, “Anomaly Network Intrusion Detection Based on Improved Self Adaptive Bayesian Algorithm”, Journal of Computers, Vol 5, pp-23-31, Jan 2010, DOI:10.4.304/jcp 5.1.

[26] Prabhjeet Kaur, Amit Kumar Sharma, Sudesh Kumar Prajapat, Madam id for intrusion detection using data mining, International Journal of Research in IT & Management Volume 2, Issue 2 (February 2012) (ISSN 2231-4334) (pg 256 – 263).

R.Venkatesan - born at Karaikal, one of the districts of Pondicherry Union Territory on 15-09-1979. Completed my post-graduation M.Sc.,(*Computer Science)* from Thanthai Hans Roever College (affiliated to Bharathidasan University, Tiruchirapalli) Perambalur, Tamil Nadu in the year 2003 & completed pre-doctoral degree M.Phil.,(*Computer Science)* from Periyar University – Salem, Tamil Nadu in the year 2009. I am pursuing PhD (*Computer Science*) in CMJ University, Shillong, Meghalaya.