Data Mining Based Database Intrusion Detection System: A Survey

***Indr Jeet Rajput, **Deshdeepak Shrivastava**
*Computer Engineering Department,
Sardar Vallabhbhai National Institute of Technology, Surat, Gujarat
India-395007
** ITM Universe, Gwalior, India

Abstract

Significant security problem for database system is unfriendly encroach a user or software. Intruder is one of the most publicized threats to security. Database is a most valuable asset of any organization or companies. This paper presents the feature of data mining based database intrusion detection system. In addition the paper gives general guidance for open research area and future direction. The objective of this survey is to give the researcher a broad overview of the work that has been done at the collaboration between intrusion detection and data mining.

Keywords: Data mining, Clustering, Intrusion Detection system, Anomaly Detection, Classification, Association.

1. INTRODUCTION

Data mining has attracted a lot of attention due to increased, generation, transmission and storage of hugs volume data and an imminent need for extracting useful information and knowledge from them. In recent year's research have started looking into the possibility of using data mining techniques in the emerging field of computer security especially in the challenging problem of intrusion detection. Intrusion is commonly defined as a set of actions that attempt to violate the integrity, confidentiality or availability of a system. Intrusion detection is the process of finding important events occurring in a computer system and analyzing them for possible presence of intrusion. Intrusion detection is a second line of defense, when all the prevention technique is compromised and an intrusion has potentially entered into the system. In general, that are two types of attacks: (i) Inside attack are the ones in which an intruder has all the privilege to access the application or the system, but it perform malicious actions. (ii) Outside attack are the ones in which the intruder does not have proper rights to access the system. Detecting inside attack is usually more difficult compare to outside attack.

2. CURRENT USED TECHNIQUES

Traditional method for intrusion detection based on signature based method. For this extensive knowledge of signature of previously known attacks is necessary. In this monitors events are matched with the signature to detect intrusion. The feature is extract from various audit database and then comparing these feature with to a set of attack signature provide by human expert for intrusion detection. The signature database has to be manually updated for each new type of intrusion that is detected.

The important boundaries with such method are that they can not detect novel attack that has a no previous signature. Another issue is that it need extensive training data for the associated technique and it may be possible they producing lots of false alarm. These all problem guide to an increasing interest in intrusion detection technique that based on data mining

Intrusion detection technique can be classified into Anomaly and misuse detection. The anomaly detection takes its decision on profile of a user normal behavior. It analyzed user's current session and compares it with the profile representing his normal behavior. An alarm raised if significant deviation is found during the comparison of session data and user profile. This type of system is useful for detection of previously unknown attacks. In contrast, misuse detection model takes decision based on comparison of user's session or commands with the rules or signature of attach previously used by attacks. The main advantage of misuse detection is that it can accurately and efficiently detect occurrence of known attacks.

3. DATA MINING TECHNIQUE

3.1 Misuse detection or Signature based: In signature based approach a signature of known attack is generated. The generated attack signature has been kept in for intrusion detection. A signature is a feature of an intruder. This approach detects only known attacks. The problem with this approach is that it is not capable to detect new attack introduced by intruder that has a no signature in database. In signature based false negative alarm rate increase. Chung et al. [1] present DEMIDS, misuse detection system for relational database systems. This method assumes that the legitimate users show some level of consistency in using the database system. If this assumption does not hold, it results in a large number of false positives. Lee et al. [2] designed a signature-based database intrusion detection system (DIDS) which detects intrusions by matching new SQL statements against a known set of transaction fingerprints. However, generating the complete set of fingerprints for all transactions and maintaining its consistency is a rigorous activity. Moreover, if any of the legitimate transaction fingerprints are missing, it can cause many false alarms. The main problem with this approach is that it is difficult to ensure that the fingerprints thus learned are indeed precise and complete.

3.2 Anomaly or Profile based: In profile based intrusion detection approach, a profile of normal user is used for intrusion detection. This approach is suitable for finding unknown attack in database. The profile of normal user is stored in database for intrusion detection. The problem with this approach is it requires more training data set. In this approach false negative alarm increased. Another problem is that significant time and effort is required for training. Zhong et al. [3] use query templates to mine user profiles. They developed an elementary transaction-level user profiles. A constrained query template is a four tuple <OP,F, T, C> where OP is the type of the SQL query, F is the set of attributes, T is the set of tables, and C is the constrained condition set. There is, however, no provision for handling various levels of granularity of access in their query template. Bertino et al. [4] proposed a database IDS that has similarity with role-based access control (RBAC) model in profile granularity. They use c-tuple (represented by query type, number of tables accessed, number of attributes accessed), m-tuple (query type, accessed table name, number of attributes accessed from individual table), and f-tuple (querytype, accessed table name, accessed attribute name) to represent a query. They build role profiles using a classifier which is then used to detect anomalous behavior. The approach Presented in [13] is query based which does not detect transaction level dependency resulting some of the database attacks may undetected. It can easily detect the attributes which are to be referred together, but it cannot detect the queries which are to be executed together. These problems resume by Rao et al. [11], this approach extracts the correlation among queries of the transaction. In this approach database log is read to extract the list of tables accessed by transaction and list of attributes read and written by transaction. Kundu et al. [5], who propose an IDS that uses inter-transactional as well as intra-transactional features for intrusion detection. It supports selection of profile and transactional feature granularity as well. In this approach a profile of normal user is generated, that can be used as a training database. It is generating profile of either user level, role level, or organization level for intrusion detection. The database schema and the purpose to achieve a meaningful task would make the user follow a particular sequence of database operations. Therefore, sequence can be a effective way of representing user profile. When a new transaction come it generate the profile using same granularity that can be used for training database development, if any deviation found in newly generated profile than that transaction to be consider as a malicious transaction. The newly generated sequence compared with the sequences stored for normal profile for identifying malicious transaction. It uses active window of a size of equal to the sequence to be taken for comparison and Alarm threshold which decided the sequence is valid or malicious. The active window and alarm threshold are the parameter to decide the efficiency of the concept. The problem with this concept is how to decide the value of this parameter.

3.3 Association rule or dependency mining: Association refers to the correlation between items in a transaction. This approach work on data dependency, in which one item is modify another item refer with this also modify. Hu et al [6] determine dependency among data items where data dependency refers to the access correlations among data items. These data dependencies are generated in the form of classification rules, i.e., before one data item is updated in the database, which other data items probably need to be read and after this data item is updated, which other data items are most likely to be updated by the same transactions. Transactions that do not follow any of the mined data dependency rules are marked as malicious transactions. The problem with this concept is that they consider only those attribute that appeared more frequently either they are sensitive or not. They treat all the attributes at the same level and of equal importance, which is not always the case in real applications. In this approach there is no concept for attribute sensitivity. Some attribute may be accessed less frequently but their modification made a more inconsistency in database. Wang et al [7] have proposed a weighted association rule mining technique in which they assign numerical weights to each item to reflect interest/intensity of the item within the transaction It is calling as weighted association rule (WAR).WAR not only improves the confidence of the rules, but also provide a mechanism to do more effective target marketing by identifying or segmenting customers based on their potential degree of loyalty or volume of purchases They first ignore the weight and find the frequent itemsets from the unweighted data and then introduce weight during rule generation. Problem with this concept is that they consider weight of attribute only rule generation. The rule is generated by using frequent itemset, if some attribute access less frequently they can't consider in frequent itemsets, that is no rule for such attribute. If this attribute is more sensitive then modification made by some intruder they can't be detected. Tao et al [8] use weighted support for discovering the significant itemsets during the frequent itemset finding phase. In this paper address the issues of discovering significant binary relationships in transaction datasets in a weighted setting Traditional model of association rule mining is adapted to handle weighted association rule mining problems where each item is allowed to have a weight In order to tackle this challenge, we made adaptation on the traditional association rule mining model under the "significant – weighted support" metric framework instead of the "large – support" framework used in previous works. In this new proposed model, the iterative generation and pruning of significant itemsets is justified by a "weighted downward closure property". Srivastava et al [9] have recently proposed the use of weighted association rule mining for speeding up web access by prefetching the URLs. These pages may be kept in a server's cache to speed up web access. Existing techniques of selecting pages to be cached do not capture a user's surfing patterns correctly. It use a Weighted Association Rule (WAR) mining technique that finds pages of the user's current interest and cache them to give faster net access. This approach captures both user's habit and interest as compared to other approaches where emphasis is only on habit. Data mining techniques can be used to mine these logs and extract association rules between the URLs requested by the users. The association rules will be of the form $X \rightarrow Y$ where X and Y are URLs. It means if a user accesses URL X then he would be accessing URL Y most likely. A. Srivastava et al [10] Database intrusion detection system is design using various approach here with we explain using data mining techniques. In this work, we have identified some of the limitations of the existing intrusion detection systems in general, and their incapability in treating database attributes at different levels of sensitivity in particular. In every database, some of the attributes are considered more sensitive to malicious modifications compared to others. Here with we explain an algorithm for finding dependencies among important data items in a relational database management system. Any transaction that does not follow these dependency rules are identified as malicious. The importance of this approach is it minimizes the number of false positive alarm. This approach generates more rules as compared to non-weighted approach. So there is a need for a mechanism to find out which of the new rules are useful for detecting malicious transactions. Such a mechanism helps in discarding redundant rules. However, the main problem with attribute dependency mining is the identification of proper support and confidence values. The attributes that are accessed infrequently may not be captured at all in the dependency rules. Use of weighted data mining algorithm reduces the problem to some extent but cannot obliterate it. The major drawback of this detection method is that the weights of attributes must be assigned manually.

3.4 Clustering Based: Cluster analysis divide data into meaningful or useful clusters in such a way that intra-cluster similarity maximized while inter-cluster similarity minimized It is a unsupervised learning technique. It is basically four types: (i) Partitioning Algorithm: in which two approach k-means and k-medoid. The k-means work for numerical data set. The main problem with k-means approach first is how to decide k values and second is determining the optimal number of clusters [12]. The k-medoid works for categorical data sets. (ii)Hierarchical Algorithm: In hierarchical clustering data are not gets clustered at ones instead stepwise procedures are followed for clustering the datasets. It resolve the problem of how to decide k value, but it has a disadvantage to specify a termination condition.(iii) Density Based Algorithm: it is based on a simple assumption that clusters are dense regions in the data space that are separated by regions of lower density. Their general idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold. M. Ester et al [13] DBSCAN, main two advantages: first, it is not sensitive to the order of data input; second, it can find clusters of arbitrary shape in spatial database with noise. The problem with DBSCAN is that it not consider many small clusters that generated after clustering that contain number of normal record, that result in high false alarm rate. (iv) Grid-based Algorithm: It is segments the data space into pieces like grid-cells. Grid-based clustering algorithms treat these grid-cells as data points, so grid-based clustering is more computationally efficient than other forms of clustering. Typical examples are STING [14].

4. SOME COMMAN IDS RESEARCH ISSUE

We have identified a number of research issues in the intrusion detection area. We can see here, developing intrusion detection system is not only about finding suitable detection algorithm but also deciding about what data to collect, how to adapt the IDS to the resources of the target system, how to test the IDS, and so on. Some of this issue is: (i) Foundation. (ii) Data collection. (iii) Detection Methods.(iv) Response. (v) Social Aspects. (vi) Operational Aspects. (vii) Testing and Evaluation (vii) IDS environment and Architecture. (ix) IDS security.

CONCLUSION

Here we have explained various data mining approach for database intrusion detection. A signature based approach suitable when pattern of attack is known. It is applicable for known attack. Association rule mining approach in which a dependency rules to be used for intrusion detection. Here we have also discussed some future issue for intrusion detection system development.

ACKNOWLEDGMENT

I will thank my teacher assistant professor Udai Pratap Rao and Dr. Dhiren R. Patel here for fervent help when I have some troubles in paper writing. I will also thank my class mates in laboratory for their concern and support both in study and life.

REFERENCES

[1] C. Y. Chung, M. Gertz and K. Levitt, “DEMIDS: A Misuse Detection System for Database Systems”, In Proceedings of the Integrity and Internal Control in Information System, Pages 159-178, 1999.

[2] 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.

[3]. Zhong, Y., Qin, X.: Database Intrusion Detection Based on User Query Frequent Itemsets Mining with Item Constraints. In: Proceeding of the 3rd international conference on information security, pp. 224–225 (2004)

[4]. Bertino, E., Terzi, E., Kamra, A., Vakali, A: “Intrusion Detection in RBAC-Administered Databases”. In: Proceedings of the 21st annual computer security applications conference (ACSAC), pp. 170–182 (2005)

[5] A. Kundu, S. Sural, A. K. Majumdar, “Database Intrusion Detection Using Sequence Alignment”. International Journal of information security volume 9, number 3, 179-191, DOI: 10.1007/s 10207-010-0102-5.

[6] Y. Hu, B. Panda, “A Data Mining Approach for Database Intrusion Detection”, Proceedings of the ACM Symposium on Applied Computing, pp. 711-716 (2004).

[7] W. Wang, J. Yang, P. S. Yu, “Efficient Mining of Weighted Association Rules”, Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 270-274 (2000).

[8] F. Tao, F. Murtagh, M. Farid, “Weighted Association Rule Mining using Weighted Support and Significance Framework”, Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 661-666 (2003).

[9] A. Srivastava, A. Bhosale, S. Sural, “Speeding up Web Access Using Weighted Association Rules”, Lecture Notes in Computer Science, Springer Verlag, Proceedings of International Conference on Pattern Recognition and Machine Intelligence (PReMI'05), pp. 660-665 (2005).

[10] Srivastava, A, Sural S., and Majumdar, AK.: “Database Intrusion Detection Using Weighted Sequence Mining”, Journal of Computers, vol. 1, no. 4 (2006)

[11] Rao, U.P., sahani, G.J., Patel, D.R., “Detection of Malicious Activity in Role Based Access Control (RBAC) Enabled Databases”. In Proceeding of Journal of Information Assurance and Security 5 (2010) 611-617.

[12] Witcha Chimphlee, et. al. “Un-supervised clustering methods for identifying rare events in anomaly detection”. In Proc. of world academy of science engg. And Tech (PWASET), Vol.8, Oct2005, pp. 253-258.

[13] M. Ester, H.-P Kriegel, J. Sander and X. Xu, “A density-based algorithm for discovering clusters in large spatial database with noise“, in Proc. of KDD-96,pp.226-231,1996.

[14]. W. Wang, J. Yang, and R. Muntz, STING: “A Statistical Information Grid Approach to Spatial Data Mining”, proceedings of 23rd International Conference on Very Large Data Bases, pp. 186-195, 1997.