International Journal of Research in Engineering, Science and Management
Volume-1, Issue-11, November-2018
www.ijresm.com | ISSN (Online): 2581-5792

An Innovative Recipe for Intrusion Detection in Relational Databases using Mean-shift Clustering and C 4.5 Algorithm

Meghana Solanki[1], Trupti Phutane[2]
[1,2]*Assistant Professor, Department of Computer Engineering, DYPCOE, Pune, India.*

***Abstract*: The seriousness of data care as well as confidentiality boosts every day. The data has become the most important resource for most companies and organizations. There are standard database security measures such as access control mechanisms, authentication and encryption technologies. They are not useful when it comes to forbidding data burglary from insiders. The security features of a Database Management System (DBMS) can be enhanced by consolidating intrusion detection mechanisms. In this paper we come up with an innovative method for recognizing intrusions in databases by applying data mining techniques such as clustering as well as classification. Experiments present that our method outruns other methods in case of high accuracy and reduced false alarm rate.**

***Keywords*: Intrusion Detection, Quiplet, User profile, Mean-shift, C4.5 algorithm.**

1. Introduction

Every institution uses database systems for storage as well as protection of confidential information. They save data not only in a well-defined but also well-organized manner. With the help of internet it is easy to access data from anyplace in the world. This easy access is very helpful to authorized users. But it become easy for intruders to make database a vulnerable. This Unauthorized access is responsible to destruction of customer's belief in the organization. Data security is the most important area for researchers. Current networks as well as operating system are not sufficient for protection of database. The database intrusions can skirt these security measures. There is triad security model: confidentiality, integrity and availability. The database is ensured by security when it satisfies above model. Unauthorized people cannot access sensitive information due to confidentiality. It put the safeguard that only the designated persons can approach the data. It remains consistent with their security level as well as access privileges. Integrity ensures the correctness as well as adherence of data. It ensures that any unauthorized person cannot alter the data. Availability ensures that only authorized users can access the desired and correct information at all times.

Violations of any of these principles causes the in security loss. Database systems have shown that insiders or outsiders of an enterprise exposes to security attacks. Insider attackers are from the organization. They are authorized person from the organization, but have a malicious intent. Outsider attackers are people who are external to the organization but have got unauthorized access to the database by escapading its security amenabilities. Outsider attacks are prevented by conventional database security techniques. The security breaches originated from both insider as well as outsider threats are identified by our proposed system. Misuse detection and anomaly detection are two well-known methods for checking intrusion. Misuse detection observes common attack patterns. They are incapable to detect malicious user actions that appear normal. Anomaly detection observes for actions that deviate from normal user behavior. There are two types of access control mechanism provided by database one is Role Based Access Control (RBAC) and other is Individual Access Control. In a RBAC database, users are grouped to different specified aspect with associated access privileges assigned to each aspect. In an individual access controlled database, rights are given to each individual. In this paper, we proposes an IDS that make use of an anomaly detection mechanism for individual access controlled databases that incorporates data mining methodologies such as clustering as well as classification. We use mean-shift for clustering and C 4.5 as the classifier for detection.

2. Literature survey

In this paper [1], [2], author gave idea about an intrusion detection of database systems. In this paper [3], an author focused on the main challenges as well as approaches used to tackle intrusions in databases. In this paper [4], an author focused on individual access control based databases. In this paper [5], [6], an author concentrated only on intrusions in RBAC based databases mechanism. In this paper [7], an author focused on insider attacks, whereas some others concentrate on outsider threats. In this paper [8], an author proposed an innovative approach for database intrusion detection is Classification of Database Transactions based on Association Rules and Cluster Analysis (CDTARCA). In CDTARCA, data dependency rules from database transactions are extracted by a rule generator. In this paper [9], an author proposed an intrusion detection approach employing a data dependency miner to

identify data correlations.

In this paper [10], an author proposed the query mining approach which uses quiplet format for representation of queries. User profiles models Normal database access behaviors that are created from database audit records. These profiles are then deployed to identify intruders. In this paper [11], an author proposed an anomaly intrusion detection model for detection of intrusions that uses an improved Apriori algorithm to set up the rules corresponding to the frequent item sets. Our intrusion detection system uses machine learning techniques like clustering as well as classification. We used Mean-shift method for clustering, since a comparison between K-means and Mean-shift method shows that Mean-shift method gives better results. For classification purpose , C 4.5 algorithm is used because comparisons between different classification algorithms shows that that C 4.5 provides more accurate results in terms of precision and recall.

3. System architecture

In our proposed system we uses Mean-shift clustering method and C 4.5 algorithm. In real world applications missing values in data are common. There are several methods that deal with this problem. In this paper we use a new version of the mean-shift clustering algorithm that concern with datasets with missing values. We make use of a weighted distance function that concern with datasets with missing values. Mean shift is a non-parametric feature-space analysis technique for locating the maxima of a density function. C4.5 is an algorithm used for generating a decision tree. Classification is done using the decision trees generated by C4.5. Due to this C4.5 is often referred to as a statistical classifier.

The main focus in our proposed method is to construct user profiles and then use these profiles for searching malicious actions that are performed in the database. The IDS undergoes the following stages.

1. Profile creation
2. Training
3. Intrusion detection

The detailed intrusion detection process is interpreted in Fig. 1. In profile creation and detection stage, quiplets are generated by converting SQL queries from database audit log. In the profile creation stage, different clusters are formed by grouping training data. Each user is aligned to a cluster. Representative cluster encompasses maximum number of records for that user. Profile for a user consists of the user id as well as the cluster into which that particular user is mapped. C 4.5 classifier is trained by making use of set of quiplets and its class information. Anomaly detection goes through two steps. In the first step, the raw SQL query is converted into a quiplet by the quiplet creator. The class associated with the user is identified by the C 4.5 classifier. In the second step, the representative cluster of the particular user is obtained from the user-cluster mapping. The class identified by the C 4.5 classifier is correlated with the representative cluster of the user if any discrepancy comes then an anomaly is raised. This output will be then given to a response file for further classification.

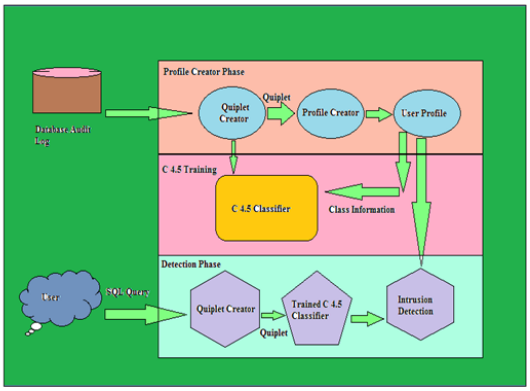

Fig. 1. Architecture for intrusion detection system

4. Data representation

At full length the IDS, quiplets are obtained from SQL queries which is a 5-ary depiction. The general depiction of a quiplet is:

(CMD, PRJREL [ ], PRJATTR [ ][ ], SELREL[ ], SELATTR[ ][ ]).

CMD is the SQL query.
SELREL[i] is a relation.
PRJATTR[i][j] is an attribute.
SELREL[i] is a relation.
SELATTR[i][j] is an attribute.

5. Algorithm

Our intrusion detection algorithm makes the use of data mining techniques such as clustering as well as classification. Profile creation is accomplished using mean-shift clustering method. Detection is carried out using class identified by C4.5 classifier. The algorithm for various steps of intrusion detection system is given below.

Profile Creation Phase
Input : Log records (R)
Output: User profile UP
For each record Ri in R
Begin
Convert Ri to quiplet
Store each quiplet to a file F along with user id
End
Cluster the data in F using Mean-Shift
For each user Ui
Begin
Let Ci be the cluster which contains maximum records for Ui
Store (Ui,Ci) into user profile UP
End
For each record Ri in F
Begin
Identify Ci for Ui from from UP
Store (Ri, Ci) into file T

End

Training and Detection Phase

Input: User profile UP, query Q, File T
Output: Searched intrusions
Convert Q to quiplet q
Train C4.5 classifier with T
Stock the C4.5 predicted result of q to Cc4.5
Identify Ci from P
if Cc4.5 = =Ci
No intrusion
else
Intrusion found.

In Mean-shift, we have used a weighted distance function that deals with datasets with missing values. Since multiple clusters are generated after clustering, we use a C 4.5 classifier for the detection phase.

6. Experimentation and result analysis

The performance of the proposed method was evaluated by conducting a number of experiments. Experiments were performed by generating database log for two databases – Printer shop and college library. Printer shop database consist of 17 tables and 6 users and the college library database consist of 9 tables and 7 users. Normal user transactions in the database are entered into the log.

Table 1
User-class mapping

User Name	Class
Manager1	1
Manager 2	1
Faculty1	3
Faculty 2	3
Applicant1	0
Applicant 2	0

Precision, recall and accuracy are the performance measurement of the system. Precision is the ratio of correctly identified intrusions by the system (TP) to total intrusions (TP+FP) given by the proposed system. Recall is the ratio of correctly identified intrusions (TP) by the system to total intrusions given by the system. Accuracy is the ratio of correctly identified intrusions as well as non-intrusions (TP+TN) to total responses from system. Table 1 shows the result of profile creation for Printer shop database. The class label used in classification is used as the cluster label. The assessment of proposed system has been done for accuracy for varying number of records. Table2 demonstrates the test results based on accuracy for the Printer shop database. A comparative analysis of both databases based on accuracy is demonstrated in Fig.2. Both these results show that accuracy increases as the number of records increase. Also, it is evident from Table 2 that the false positive rate reduces with the increase in number of records. It is clear from the result that our system outperforms both these methods in terms of precision and recall.

Table 2
Test result for ID process in printer shopl database

Number of test records	False Positive Rate (%)	False Negative Rate (%)	Precision	Recall	Accuracy
60	22	11	89.05	91	88
110	14.3	9.11	93.45	92.55	92
160	9.68	8.17	93.74	94.65	94.89
210	9.56	5.46	95.62	96.52	96

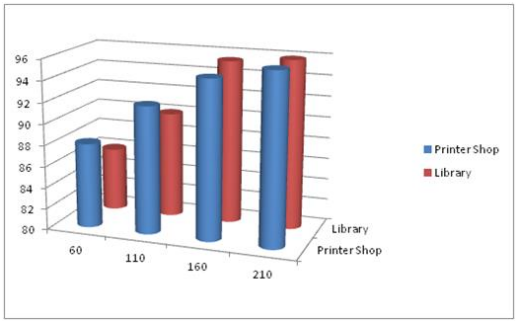

Fig. 2. Comparison of accuracy for two databases

7. Guidelines

We recommended an innovative database intrusion technique in case of relational databases by making use of data mining techniques such as Mean-Shift clustering method and C 4.5 classification algorithm. Anomalies are effectively detected by the proposed system .They can be applied for detecting intrusions with higher accuracy as well as reduced false alarm rate. An experiment demonstrates that our system works productively in search of intrusions. The results show that accuracy and performance of the system increases as there is increase in records.

References

[1] Y. Hu and B. Panda, "A Data Mining Approach for Database Intrusion Detection," *ACM Symp. Appl. Comput.*, pp. 711–716, 2004.
[2] M. Doroudian and H. R. Shahriari, "A hybrid approach for database intrusion detection at transaction and inter-transaction levels," *2014* 6th Conf. Inf. Knowl. Technol. IKT 2014, no. Ikt, pp. 1–6, 2014.
[3] R. J. Santos, J. Bernardino, and M. Vieira, "Approaches and Challenges in Database Intrusion Detection," *ACM SIGMOD Rec.*, vol. 43, no. 3, pp. 36–47, 2014.
[4] Z. Yanyan and Y. Yuan, "Study of database intrusion detection based on improved association rule algorithm," *Proc. - 2010 3rd* IEEE Int. Conf. Comput. Sci. Inf. Technol. ICCSIT 2010, vol. 4, pp. 673–676, 2010.
[5] S. M. Darwish, S. K. Guirguis, and M. M. Ghozlan, "Intrusion detection in role administrated database: Transaction-based approach," Proc. - 2013 8th Int. Conf. Comput. Eng. Syst. ICCES 2013, pp. 73–79, 2013.
[6] U. P. Rao, G. J. Sahani, and D. R. Patel, "Machine learning proposed approach for detecting database intrusions in RBAC enabled databases," 2010 2nd Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2010, pp. 1–4, 2010.
[7] H. Q. and S. U. Sunu Mathew, Michalis Petropoulos, "A Data- Centric Approach to Insider Attack Detection in Database Systems," *RAID 2010*, pp. 382–401, 2010.
[8] I. Singh, V. Darbari, L. Kejriwal, and A. Agarwal, "Conditional adherence based classification of transactions for database intrusion

detection and prevention," 2016 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2016, pp. 42–49, 2016.

[9] Y. Hu and B. Panda, "A Data Mining Approach for Database Intrusion Detection," *ACM Symp. Appl. Comput.*, pp. 711–716, 2004.

[10] A. Kamra, E. Terzi, and E. Bertino, "Detecting anomalous access patterns in relational databases," *VLDB J.*, vol. 17, no. 5, pp. 1063–1077, 2008.

[11] W. Gongxing and H. Yimin, "Design of a New Intrusion Detection System Based on Database," *2009 Int. Conf. Signal Process. Syst.*, pp. 814–817, 2009.