Write Up Of Project

In today's highly connected world where more and more organizations are setting up online and connecting to their customers over the internet, the data generated and stored by these organizations is growing at an exponential rate. The data collected by the organization is critical for its day to day operations and future decision-making policies, and the data stored by the organization in various structures like database management systems is a crucial asset for the organization. Therefore, the organization must take steps to ensure the integrity, confidence, and availability of the systems used to store the data.

This issue is further exasperated by the increase in web applications and information systems that increases the risk of exposure of the databases, therefore, database security is more crucial today than ever before. Through this project, we are trying to develop a system to ensure database security for various organizations, like banks and hospitals to ensure the confidentiality of banking details, patient information, etc.

It is also vital to address the increasingly important issue of insider threats to the organization. In general insider threats are more menacing and dangerous than outside threats like hacking, malware, etc. This is because, an insider already has the proper authorization for access to the database and also knows the nuances of the security systems put in place due to which, it is increasingly difficult to identify malicious transactions made by an insider in comparison to outside threats. Therefore, through this method, we intend to develop novel and effective techniques to identify malicious transactions from both inside and outside threats.

In this project, we utilize various data mining techniques like Sequential Pattern Mining(SPM) and Association Rule Mining(ARM) to develop data dependencies among the various attributes in the database. These data dependencies can then be used to check if a new transaction follows them, depending on which the transaction may be classified as either a malicious or safe transaction. In addition to this, we also make use of various anomaly detection techniques like statistical modeling, proximity-based detection, clustering, etc. to find patterns in the transactions that can be used to identify whether the transaction is malicious or safe. In the end, we then combine the result obtained from the two systems described above to finally assign an anomaly score to the transaction and label it as either malicious or safe.

Sequential Pattern Mining or SPM is a topic of data mining concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence. In this, all the sub-sequences that satisfy the minimum support in a set of sequence patterns are selected. Various SPM algorithms that can be used are Generalized Sequential Pattern Mining, SPADE algorithm, SPAN algorithm, etc.

Association Rule Mining or ARM is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measure of interestingness. It is meant to find frequent patterns, correlations, associations in data sets that are present in the database. Various ARM algorithms that are used are the Apriori algorithm, AprioriTid Algorithm, etc.

The SPM and ARM algorithms together can then be used to identify existing patterns in the operations carried out in any given transaction. This can then be used to define data dependencies that can be used to check if a given transaction is malicious. Also, since each attribute in any given database may have a varying degree of importance or sensitivity, we can assign each of these attributes different weights. For example, in a banking system, the account number and debit card number are more important and sensitive in comparison to the names and other personal details of the account holder. Thus it is more important to track the malicious transactions according to their sensitivities.

With just the above system in place, the system would need to be updated from time to time, to update the various data dependencies. This is a trivial task, and if a malicious transaction is run that does not oppose a dependency, then it would be allowed without raising any alarms. Therefore, we also need a system in place that can identify malicious transactions that have not been seen before and those malicious transactions that cannot be identified by the first system. Therefore, for this purpose, we employ an anomaly detection system, which supports the detection of new attacks.

The anomaly detection is done with the help of various statistical and probabilistic models and machine learning techniques like proximity-based techniques, clustering, and other machine learning algorithms. Clustering analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other

groups. In clustering-based Anomaly Detection, small clusters or even individual data points that lie far from the other clusters in low-density regions are classified as outliers and given an outlier score. If an outlier has an outlier score greater than the predetermined threshold, then it is considered as anomalous.

Various clustering algorithms can be used for clustering like the K-Means Clustering algorithm, Fuzzy C-Mean Clustering, Agglomerative Hierarchical Clustering, etc. Each of these clustering algorithms has an advantage or disadvantage when compared to each other.

Finally, the outputs from the above two systems are combined together with weight factors to make the final decision about the malignity of the transactions.

We are currently working to use the improved versions of studied algorithms and integrate different concepts mentioned to develop robust IDS that can be deployed in organizations to safeguard its database.