RUTGERS UNIVERSITY
PhD in Management Program

EXPERT SYSTEMS
Prof. Glenn Shafer

Detecting Inference Attacks Using Association Rules

Abstract: An Inference attack occurs when a user is able to infer some information about a database without directly accessing them. Data mining techniques help to discover unknown information from a database using known information. This paper discusses how Association Rule mining can be used to detect unauthorized inferencing from the database.

Sangeetha Raman

13, December 2001.

Detecting Inference Attacks using Association Rules

1. Introduction

Regulating and monitoring user access to database has been one of the important concerns of the database security community. While data mining techniques are useful in discovering information from a database, such techniques could expose the data to security threats if the data that is mined reveals information that is considered ‘confidential’. Unauthorized inference of data occurs when the information that is mined reveals ‘confidential’ details about the organization, or when generalizations about the overall database can be made based on the individual data.

“An inference attack occurs in multilevel secure database when a low level user is able to infer sensitive information through common knowledge and authorized query responses “ [2]. Although the actions taken to prevent users from making unauthorized inferences from a database are critical to database security, the focus of this paper is only on the first step in handling unauthorized inferencing – detecting when users could make such unauthorized inferences. The focus of this paper is on how Association Rules can be used for this purpose.

2. What is an inference attack?

Inference problems are security concerns that arise when users deduce sensitive information about the database from relatively trivial information. Inference problems differ from other security problems in that it is not an issue of unauthorized access to data or leakage of information. Rather, unauthorized inferencing is the result of the nature of the information and the semantics of the

application itself. Existing approaches to inference detection focus on analyzing functional dependencies (i.e., the relationship between the attributes) in the database structure. It is also possible to detect unauthorized inferences at the data level itself.

"Inference or the inference problem is that of users deducing (or inferring) higher level information based upon lower, visible data" --- [Morgenstern 88]

Inferencing is the activity of inferring new (previously unknown) information from known information. This becomes a problem when inferring extends to deducing unauthorized information from authorized access of data. The problem of inference detection is quite complicated not only because of the different ways in which inferencing can be done but also because of the variety of information from which inferencing is done.

Inference problems can be studied in terms of inference channels [Garvey 94]. Some of the different types of inference channels are:

1) *Deductive Inference Channel:* This channel is the most restrictive one, requiring a formal proof of the deductive methodology showing that HIGH data can be derived from LOW data.
2) *Abductive Inference Channel:* In this channel, it is sufficient to show that a deductive proof can be completed assuming certain LOW-level data
3) *Probabilistic Inference Channel:* This channel of inferencing arises when the probability of inferencing unauthorized information is greater than an acceptable level.

3. Inference Rules:

The goal of inference detection is to determine if a user can indirectly access data using multiple (and relatively trivial) queries. In order to handle the problem of unauthorized inferencing, we need to understand the different strategies that can be used in making inferences. Inferencing, whether authorized or unauthorized, involves the application of certain rules [11]. Understanding these

rules is the first stage in developing an effective inference detection system. Some of the rules used in inferencing are:

1. Subsume Rule: This rule handles the type of inferencing that can occur when the result of one query and the result of another query together correspond to the same tuple (row) in the database.
2. Overlapping Rule: This rule helps to identify the common values of a set of queries when only some of the values returned by a query match some of the values of another query.
3. Complementary Rule: This rule helps to identify the corresponding return values of one query by taking the difference between two sets of queries.
4. Functional Dependency Rule: This refers to inferencing at the structural level based on the relationship between the attributes of a database.

Some methods of inferencing

It is almost impossible to develop an inference detection system that is capable of detecting and controlling all types of inference attacks. The following list [12] provides some of the strategies that can be used to make inferences from databases.

- *Inference by Deductive Reasoning:* This refers to inferencing based on established rules in classical or non-classical logic deduction. Classical deduction is based on the basic logic rules while non-classical deduction may be done using probabilistic reasoning, fuzzy reasoning, etc.
- *Inference by Inductive Reasoning:* This refers to inferencing when established rules are used to develop hypothesis from observed examples
- *Inference by Analogical Reasoning:* This method involves obtaining new information about a variable given the properties of another variable.
- *Inference by Heuristic Reasoning:* Heuristics may also be used to deduce information.

- *Inference by Semantic Association:* This involves determining the association between entities based on the information given about the entities themselves.
- *Inferred Existence:* The existence of certain entities can be inferred based on the available information
- *Statistical Inference:* This inferencing method involves determining information about an individual entity using statistics on various other entities.

4. Inference Information

Another problem in inference detection concerns the vast variety of information from which inferences can be made. Since users can obtain information from a variety of sources, it becomes very difficult to track the specific path that is used in inferencing. This makes the task of developing a comprehensive inference detection system a much more difficult one. A list [12] of the possible types of information that can be used to make inferences is given below:

(1) Information that is stored in the database
(2) The design of the database
(3) The relationship between the different attributes of the database
(4) Statistical data derived from the database
(5) The existence or absence of data
(6) The changing values of the data
(7) Specialized information about the database that is not available from the database but known to certain persons in the organization
(8) Common knowledge and Common sense

5. Association Rule in Inference Detection

The Knowledge Discovery and Data Mining Technology has been widely used to enforce database security. While Data mining techniques like classification, clustering, characterization and association mining offer valuable tools to learn about the data, they also expose the data to security risks. The first step in detecting inference attacks is to analyze whether the results of the user queries reveal any information that is sensitive. Predicting the overall effect from the set of queries could help in determining whether the user could make an unauthorized inference from the database. Prediction involves generating patterns in the queries issued by the user. Data mining algorithms play a major role in determining whether any relationship exists between the attributes, and if so, the pattern of those relationships and how they can be interpreted. Although a naïve prediction could be made about the data, such a prediction cannot determine the certainty levels calculated through Association Rules.

Overview of Association Rule

'An association rule is a simple probabilistic statement about the co-occurrence of certain events in a database ' [6]. An association rule is of the form $X \rightarrow Y$ [c, s] [6] where X and Y are item sets and $X \wedge Y = \varnothing$, s = support (X UY) and c = {support (XUY) / support (X)} is the confidence.

Association rule mining helps in detecting inferencing by providing information about possible associations between the attributes and areas where the database is exposed to inference risks. An effective inference detection system should be able to ascertain that certain specific information can be inferred from the database, and if so, what information was used to perform the inference [4].

Since inferencing does not involve direct breach of the database security system, detecting instances of unauthorized inferencing is a very difficult task. There is no single security system that is capable of detecting all types of inferencing. Most of the inference detection systems that are used these days are developed on a case-by-case basis. Association rules are useful in determining possible relations between attributes and in predicting how inferences can be made from the attribute-relation structure of the database.

The pattern of user queries can be used to predict the overall result of the specific query or a set of queries. For example, let us consider a scenario where certain values of an attribute ‘A’ in a database are considered sensitive and our goal is to prevent certain users from inferring those values. We can determine all the possible associations between ‘A’ and the other attributes and the support and accuracy levels for each such association using association rules. The set of associations that reveal the sensitive data can be categorized as sensitive and security mechanisms can be enforced to track the pattern of querying by users and rejecting the queries that follow the pattern categorized as sensitive.

An example of how association rules can be used in detecting unauthorized inferencing from databases is given below.

Consider the following simple database of a Bank:

Acct_no	Name	Address	Preferred	Balance	Acct_type
1	A	123, High Rd	Low	500	Checking
2	B	145, Low Ave	Mod	1000	Savings
3	B	145, Low Ave	Mod	1500	Checking
4	C	189, Park Dr	High	5000	Savings
5	D	200, Main St	High	4500	Savings
6	E	30, Plain Rd	Low	800	Checking
7	E	30, Plain Rd	Low	300	Savings
8	F	140,Wood Ave	Mod	1000	Checking

9	G	56, River Rd	High	600	Savings
10	G	56, River Rd	High	4800	Checking

Let us assume that the security policy of the Bank is to prevent junior tellers from inferring the details about whether a customer is highly preferred, moderately preferred or lightly ('low') preferred (i.e., the number and names of customers in each category). It is possible to enforce a security system for this purpose. But such a system can only prevent the junior tellers from obtaining the data by directly querying the database. The security system cannot prevent the junior tellers from inferring the data by issuing indirect queries. Association rules could be very useful in such situations. They could be used to identify all the possible ways, by which 'non-confidential' data can reveal 'confidential' data, i.e., identifying the different paths that can reveal such data.

Some Association Rules generated from the above data set:

Rule 1: IF (Acct_type = 'Savings') $\wedge$ (Balance $\geq$ 2000) THEN
(Preferred = 'High') [50%]

Rule 2: IF (Acct_type = 'Savings') $\wedge$ (Balance < 2000) THEN
(Preferred = 'Med') [33%]

Rule 3: IF (Acct_type = 'Savings') $\wedge$ (Balance < 2000) THEN
(Preferred = 'Low') [33%]

Rule 4: IF (Acct_type = 'Checking') $\wedge$ (Balance $\geq$ 2000) THEN
(Preferred = 'High') [100%]

Rule 5: IF (Acct_type = 'Checking') $\wedge$ (Balance < 2000) THEN
(Preferred = 'Med') [50%]

Rule 6: IF (Acct_type = 'Checking') $\wedge$ (Balance < 2000) THEN
(Preferred = 'Low') [50%]

A junior teller could infer the number of highly preferred customers by using classification and association rules without directly querying the database. In the above example, a junior teller could infer the details by issuing multiple queries on the database. For instance, a junior teller could issue a query to retrieve the names of the customers who belong to the ‘med’ or ‘low’ preferred class and then subsequently issue a query to retrieve the names of customers whose balance is above $ 500. Thus our junior teller will be able to obtain the number of the highly preferred customers by combining the results of the two queries with some common knowledge that he/she has about the database.

One way to detect unauthorized inferencing is to mine the database first and generate all the possible Association Rules that exist in it. The rules so generated can be used as a basis for generating all possible sets of patterns that could reveal ‘confidential’ information. This set could then be used for comparison with the users’ activities on the database. A security system that would signal whenever a querying pattern follows this set can be enforced to identify unauthorized inferencing.

Preventing users from making such unauthorized inferences from the database can be a difficult task because of two primary reasons. Firstly, it is difficult to predict what information the user already has about the database and secondly, it is difficult to predict the ‘path’ that the user would follow in order to make the inferences. While the second problem can be handled through data mining techniques, the first problem depends upon the individual organization’s confidentiality policy.

One way to prevent unauthorized inferencing of data is to limit user access to the database. If the users are prevented from obtaining a large sample of the database, they can be dissuaded from developing association rules with high certainty levels. Another approach is to alter the data so that individual details are

accurate but overall generalizations from the database are inaccurate. Including ‘dummy’ data in the results returned by the query controls unauthorized generalizations about the database by ensuring that the conclusions drawn about the database are incorrect.

Clifton et al [1] suggest that data mining techniques can be used to determine if the data is mineable. Such techniques could then be used to develop ‘profiles’ of hard-to-mine data. Using such profiles, we can build systems to ensure that the results mined are either ‘non-interesting’ or incorrect because of ‘planted’ values [1].

While enforcing restrictions on user access may prevent unauthorized inferencing, too much restriction could seriously hinder the normal functioning of the organization.

Conclusion:

Data mining offers valuable tools for knowledge discovery from databases. One such tool is Association Rule mining. Association rules reveal important associations among the data that is critical to data management. While association rules cannot be used to control inference attacks, they offer a useful technique to predict the extent and accuracy with which inferences can be made about the information contained in the database. Although the actual process of preventing unauthorized inferencing from the database is a very difficult task, understanding the inferencing pattern could prove to be very effective in handling the problem.

Preventing unauthorized inferencing could be a challenge to the database security community. Since inferencing from the database is made with a certain level of confidence and support, we can ensure that the rules 'discovered' to make such inferences are either incorrect or are derived with higher confidence and support than the actual levels [8].

There are some Association Rules that require large reruns on the data even to form the rules. This feature can be exploited to deter users from making unauthorized inferences by ensuring that the volume of data retrieved is too large for them to develop rules.

Hence, Data Mining techniques are very useful in the area of inference detection. Some of the Data Mining algorithms like classification rule mining can also be combined with association rule mining to develop stronger monitoring systems. This objective can be achieved by classifying the association rules to determine specific patterns in them that could reveal new ways of making (unauthorized) inferences from the database.

References:

[1] Clifton, C. and Marks, D. (1996): *Security and Privacy Implications of data mining*

[2] R. Agarwal, T. Imielinski and A. Swami: *Mining Association Rules between sets of items in large databases*

[3] Salvador Mardujano: *Data Suppression, Concealing Controls and other security trends*

[4] Harry S. Delugach and Thomas H. Hinke: *Wizard: A Database Inference Analysis and Detection System*

[5] T. H. Hinke, D. S. Delugach and R. Wolf: *A framework for Inference-directed Data Mining*

[6] Hand, Mannila and Smyth: *Draft of Principles of Data Mining*

[7] Tom Johnsten and Vijay V. Raghavan: *Impact of Decision-Region Based Classification Mining Algorithm on Database Security*

[8] Chris Clifton: *Protecting Against Data Mining through Samples*

[9] Wenke Lee and Salvatore J. Stolfo: *Data Mining Approaches for Intrusion Detection*

[10] Wenke Lee, Salvatore J. Stolfo and Kui W. Mok: *A Data Mining Framework for Building Intrusion Detection Models*

[11] Raymond W. Yip and Karl N. Levitt: *Data Level Inference Detection in Database Systems*

[12] *NCSE Technical Report –005 Volume 1/5 May 1996*