Chapter 13

ON SECURITY AND PRIVACY RISKS IN ASSOCIATION MINING ALGORITHMS

Tom Johnsten, Vijay Raghavan and Kevin Hill

Abstract Association mining algorithms can be used to discover sensitive knowledge from non-protected data that are voluntarily released for mining purposes. In this paper, we present a risk assessment procedure designed to control the security and privacy threats presented by association mining algorithms. The assessment procedure is based upon a new data analysis process that we introduce called Knowledge Hiding in Databases (KHD). The specific contributions of this paper include a description of the KHD process, and the development and implementation of a risk assessment procedure for association mining algorithms.

Keywords: Knowledge hiding in databases, association mining, security, privacy

1. Introduction

Nowadays companies and organizations frequently use KDD technology to analyze their stored data to discover valuable patterns or rules that can help them maintain their competitive edge. Recently, researchers within the information security community have begun to examine the impact of this technology on database security [1,3-9,12]. The primary focus of this work has been on the security threat presented by classification mining. In contrast, the work reported on in this paper is concerned with the security and privacy threat presented by association mining [7,12].

To illustrate the security threat presented by association mining, consider the data shown in Tables 1 and 2. Suppose that each transaction in Table 1 corresponds to items purchased at a local health store. The *Cust-Id* data is obtained from loyalty cards that the health store offers to its customers. To acquire a loyalty card a customer fills out a form that includes such personal information as name and home address. A loyalty card benefits customers by making them eligible for a discount on purchased items. The security threat presented in this example is the extent to which the collected data facilitates

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: [10.1007/978-0-387-35697-6_26](https://doi.org/10.1007/978-0-387-35697-6_26)

E. Gudes et al. (eds.), *Research Directions in Data and Applications Security*

© IFIP International Federation for Information Processing 2003

Table 1. Health store transaction data.

CUST-ID	St-Johns-Wort	Ginkgo-Biloba	Cats-Claw	Maitake	LAR_PUR
001	1	1	0	1	1
011	0	1	0	1	0
023	0	1	1	0	0
004	1	1	1	0	0
001	1	0	0	0	1
006	1	0	1	0	0
:	:	:	:	:	:
:	:	:	:	:	:

Table 2. Health store customer data.

CUST-ID	Name	Address
001	Chris Bennet	701 S. Randolph
004	Linda Jones	100 W. Jefferson
006	Steve Anderson	808 E. Murray
011	Anne Richardson	107 N. Chandler
023	Jack VanBrooker	218 W. Jefferson
:	:	:
:	:	:

the identification of individuals with specific health conditions. For example, a store employee with the appropriate background knowledge could construct a concept hierarchy to identify transactions from Table 1 associated with a particular health condition. The employee could then apply an association mining algorithm, along with the hierarchy, to identify individual customers from Table 2 associated with the targeted health condition.

In [6], Clifton and Marks provide an alternative example to illustrate the security threat presented by association mining. Their example is based on a supermarket that releases its customer-transaction data to a wholesale distributor. The distributor contends that the data will allow them to track inventory and reduce warehouse costs. However, the distributor, through the application of an association mining algorithm, determines that customers who purchase their competitor's product X also typically purchase product Y. As a result, the distributor offers customers a discount on product Y when purchased in conjunction with their own equivalent for product X. The security threat is the extent to which the discovered information may result in unfair competition among the distributors.

The two examples presented above motivate the need to develop counter-measures to address the security and privacy threats related to association mining. To that end, the rest of this paper is organized as follows. Section 2 presents a methodology to analyze the impact of KDD technology on database security. The methodology consists of six steps that make up a new data anal-

ysis process that we call Knowledge Hiding in Databases (KHD). The goal of KHD, in contrast to KDD, is the non-trivial hiding of potentially sensitive knowledge in data. Section 3 outlines a risk assessment procedure designed to control the security and privacy threats presented by association mining. The proposed procedure is based upon the KHD methodology presented in Section 2. Section 4 describes a security assessment tool developed as part of this work. The developed tool, referred to as the Association Mining Risk Assessment Tool (AMRAT), allows for the implementation of the assessment procedure presented in Section 3. Finally, Section 5 summarizes the paper and discusses some future related research projects.

2. KHD Methodology

As mentioned in the introduction, the goal of Knowledge Hiding in Databases (KHD) is the non-trivial hiding of potentially sensitive knowledge in data. We define the term 'non-trivial' to imply that knowledge is concealed in a manner that maximizes the amount of released data and maintains, to the greatest extent possible, the integrity of the data. In this context, data integrity is defined as the ability to obtain accurate knowledge from parts of the data that are legitimately available. The inputs into the KHD process are a set of security constraints that must be satisfied and a collection of data D ; and, the output is a collection of data D' , derived from D , that satisfies the given security constraints. In general, KHD is an iterative process consisting of the following six steps:

- Identification of Sensitive Knowledge
- Identification of Appropriate Classes of Data Mining Algorithms
- Formulation of Security Policies
- Risk Assessment
- Sanitization of Data
- Report Generation

The first step is to identify the sensitive knowledge that needs to be concealed within the data. For example the statement, "store employees may not have knowledge of individual customers suffering from sensitive health conditions", represents knowledge, or information, that an organization may wish to conceal. The identification of sensitive knowledge is an important task since the resulting analysis will only be as complete as the identified knowledge. In general, such knowledge is derived from an organization's stated security policies and its collected data. The second step is to match the knowledge identified in the first step with classes of data mining algorithms that are capable of

discovering it. For example, the previous inference related to "sensitive health conditions" could be discovered through the application of an association mining algorithm. Of course, if there does not exist a data mining algorithm that is capable of discovering some piece of knowledge K , then K is concealed within the data.

The third step is to translate the sensitive knowledge identified in the first step into security policies that can be evaluated against the collected data. The structure of the policies will depend, in large part, on the classes of data mining algorithms identified in the second step. The fourth step is to evaluate the formulated policies against the collected data to determine if there are policy violations. If it is determined that there will be violations, then it is necessary to sanitize the data in order to conceal, or hide, the sensitive knowledge. In general, modifying a collection of data involves altering existing data, concealing existing data, or introducing additional (spurious) data. The specific manner in which data are modified is dependent upon several factors including, the data mining algorithms identified in the second step, the need to maximize the amount of released data, and the need to maintain the integrity of the data. A subtle, but significant, issue is that in some instances the act of modifying the data may affect the fundamental behavior of the targeted data mining algorithms [9]. The sixth, and final step, is the generation of a report describing the results of the KHD process. The report should include detailed information about the integrity of the sanitized data.

3. KHD Process: Association Mining

This section of the paper examines the steps of the KHD process in the context of association mining algorithms.

3.1 Identification of Sensitive Knowledge

Association mining can be used to discover sensitive knowledge that is expressed as an association among stored items. For example, the discovery of associations among the items in Table 1 may lead to the disclosure of customers suffering from sensitive health conditions. In order to identify such knowledge, or information, we propose the use of "fault-trees" along with an organization's security policies.

The construction of a fault-tree requires the identification of a top-event and the causal events related to the top-event [11]. In the current context, a top-event represents a specific piece of knowledge that an organization wishes to conceal within their collected data; while, a causal event represents either the discovery of a piece of knowledge or the occurrence of transactions that include a subset of items. Causal events that represent the discovery of a piece of knowledge are specified as non-terminal nodes and those that identify specific

Figure 1. Example fault-tree.

transactions are represented as terminal nodes. An example of a constructed fault-tree is shown in Figure 1. This fault-tree corresponds to the health store example in Table 1, where the sensitive knowledge includes transactions involving the purchase of items related to depression or AIDS, or transactions that include items frequently occurring in purchases of \$25.00 or more. There are three terminal nodes in the fault-tree. Individually, these nodes express an association between a *CUST_ID* item and *depression* items, a *CUST_ID* item and *AIDS* items, and a subset of stored items and a *LAR_PUR* item. The terms *depression* and *AIDS* represent concepts that are defined in terms of items stored in the database.

3.2 Identification of Data Mining Algorithms

As stated above, the second step in the KHD process is to match the knowledge identified in the first step with data mining algorithms that are capable of discovering it. Our focus here is on the sensitive knowledge that can be discovered through association mining. For instance, the knowledge defined in Figure 1 is relevant since it may be obtained through the application of an association mining algorithm.

3.3 Formulation of Security Policies

This step requires transforming the identified sensitive knowledge into an appropriate set of security policies that can be evaluated against the collected data. To facilitate such a transformation, we propose the use of a predefined collection of templates. A template represents one or more association rules and has three user-defined threshold values that correspond to a measured level of support, confidence and improvement. The measurements of support, confidence and improvement are typically used to evaluate the accuracy and strength of a discovered association rule [2,10]. The support and confidence of an association rule, $X \rightarrow Y$ (where X and Y are sets of items), are defined as the percentage of transactions in the database that contain both X and Y , and the percentage of transactions containing X that contain Y , respectively. The measurement of improvement is defined as the joint probability of X and Y divided by the product of the probability of X and the probability of Y [2]. If the improvement is greater than one, the rule is better at predicting the result than random chance.

The user-defined threshold values reflect the sensitivity of the knowledge represented by the template. A security policy defined in terms of a template is satisfied if and only if the calculated level of support, confidence and improvement are all below their corresponding threshold value. Our concept of a "template" is analogous to the concept of an "inclusive" template as used by association mining algorithms to direct the knowledge discovery process [10]. However, the proposed templates represent knowledge for which the reported accuracy should be relatively low (in spite of the robustness of the data mining algorithm).

As a part of this work, we have designed twelve distinct templates for specifying association-based security policies. The templates are listed in Figure 2. It is our contention that most, if not all, association-based security policies can be specified in terms of these twelve templates. The templates are defined in terms of a specific item, any item, any subset of items and a specific concept. In particular, a "specific" item is a single named item, an "any" item is a placeholder for an arbitrary single item, an "any subset" item is a placeholder for an arbitrary subset of items, and a "specific" concept is a named subset of items. Several of the templates, including template types -2, -3, -4, -5, -8, -9, -10 and -11, represent multiple association rules. The templates corresponding to the fault-tree in Figure 1: "TYPE-6: CUST_ID $\rightarrow$ Depression", "TYPE-6: CUST_ID $\rightarrow$ AIDS, and "TYPE-5: Any Subset of Items $\rightarrow$ LAR_PUR. The first and second templates identify customers associated with transactions involving the purchase of items related to Depression and AIDS, respectively. The third template identifies subsets of items that occur with transactions involving the LAR_PUR item.

TYPE-1:	Specific Item $\rightarrow$ Specific Item
TYPE-2:	Specific Item $\rightarrow$ Any Item
TYPE-3:	Any Item $\rightarrow$ Specific Item
TYPE-4:	Specific Item $\rightarrow$ Any Subset of Items
TYPE-5:	Any Subset of Items $\rightarrow$ Specific Item
TYPE-6:	Specific Item $\rightarrow$ Specific Concept
TYPE-7:	Specific Concept $\rightarrow$ Specific Item
TYPE-8:	Any Item $\rightarrow$ Specific Concept
TYPE-9:	Specific Concept $\rightarrow$ Any Item
TYPE-10:	Any Subset of Items $\rightarrow$ Specific Concept
TYPE-11:	Specific Concept $\rightarrow$ Any Subset of Items
TYPE-12:	Specific Concept $\rightarrow$ Specific Concept

Figure 2. Security templates for association mining.

3.4 Risk Assessment

The risk assessment step requires the evaluation of security policies to determine if there are policy violations. Our procedure for evaluating policies defined in terms of templates for association mining consist of three general steps. First, each template is expanded into one or more corresponding association rules. Second, each rule is evaluated against the collected data using the measures of support, confidence, and improvement. Third, each rule is recorded along with an indication as to whether there is a violation of security policy. Section 4 describes the implementation of this procedure in the context of AMRAT.

3.5 Sanitization of Data

The process of sanitizing data requires reducing an association rule's reported level of accuracy and strength to ensure that there is no violation of policy. We present two different approaches for sanitizing a collection of data.

One approach is to remove items from the database to ensure that the association rules that violate policy no longer exist. In order to maximize the amount of available data, we introduce the concept of a Minimum Coverage Item Set (MCIS). We define a MCIS as follows. Given a set of association rules A , a MCIS is a minimum set of items in which at least one of the items in the set is included in each rule $a \in A$. This approach to sanitizing data guarantees that all rules that violate their security policy have a zero level of support, confidence and improvement with respect to the sanitized data. In addition, this approach maintains the integrity of the data since the values of individual items are not altered. Therefore, any "non-sensitive" association rule discovered from the sanitized data will have the same measured level of support, confidence, and improvement as with respect to the unsanitized data.

An alternative approach is to selectively modify certain item values [7,12]. We have developed three separate procedures designed to reduce an association rule's level of support, confidence and improvement to ensure that there is no violation of policy. In each procedure, the primary step is the determination of the number of item values, N , that must be altered in order to achieve the desired outcome. Once the value of N is determined, the next step is to select an item from the rule's right-hand side for which to alter N of its values. The transactions that are altered must not only include the selected right-hand side item, but must also satisfy the logical conjunction of the rule's left-hand and right-hand sides. The process of altering item values with respect to a specific rule may cause a violation of security policy with respect to other rules. As a result, all association rules must be iteratively evaluated until there is no reported policy violation.

The advantage of the second approach is that it maximizes the amount of available data, although it does not ensure the integrity of the data. In general, the modification of item values will alter the discovery of non-sensitive association rules. However, it is possible to determine the impact that the modification of an item's values has on a rule's measured support and confidence. Specifically, for a given rule $X \rightarrow Y$ discovered from the sanitized data there are four possible cases to consider and they are outlined in Figure 3.

Case-I addresses the situation in which data integrity is maintained. In particular, if $X \rightarrow Y$ contains no items whose values have been modified, then $X \rightarrow Y$ will have the same measured level of support and confidence as with respect to the unsanitized data. Case-II addresses the situation in which items whose values have been modified are included as part of the rule's left-hand side. The symbol $\#(X \& Y)$ denotes the number of transactions in the sanitized data that satisfy the logical expression $(X \& Y)$, and $P_MAX(X)$ is the maximum number of altered values over the items $x \in X$. To illustrate, suppose the non-sensitive association rule, $x_1 \& x_2 \rightarrow y$, is discovered in the sanitized data with a measured level of support and confidence of $(30/100 = 30\%)$ and $(30/50 = 60\%)$, respectively. If $P_MAX(X)$ is 6, then the actual level of support and confidence is within the interval $[30.0\%, 36.0\%]$ and $[53.5\%, 64.2\%]$, respectively. Case-III addresses the situation in which items whose values have been modified are included as part of the rule's right-hand side. $P_MAX(Y)$ is the maximum number of altered values over the items $y \in Y$. Consider again the non-sensitive association rule, $x_1 \& x_2 \rightarrow y$, with support equal to 30% and confidence equal to 60%. If $P_MAX(Y)$ is 10, then the actual level of support and confidence is within the interval $[30.0\%, 40.0\%]$ and $[60.0\%, 80.0\%]$, respectively. Finally, Case-IV addresses the situation in which items whose values have been modified are included as part of the rule's left-hand side and right-hand side. $P_MAX(X, Y)$ is the maximum number of altered values over the items $x \in X$ and $y \in Y$. The

CASE-I:

Support: data integrity maintained

Confidence: data integrity maintained

CASE-II:

Support: $[(\#(X \& Y) / \text{num_transactions}) * 100, ((\#(X \& Y) + P_MAX(X)) / \text{num_transactions}) * 100]$

Confidence: $[(\#(X \& Y) / (\#(X) + P_MAX(X))) * 100, ((\#(X \& Y) + P_MAX(X)) / (\#(X) + P_MAX(X))) * 100]$

CASE-III:

Support: $[(\#(X \& Y) / \text{num_transactions}) * 100, ((\#(X \& Y) + P_MAX(Y)) / \text{num_transactions}) * 100]$

Confidence: $[(\#(X \& Y) / \#(X)) * 100, ((\#(X \& Y) + P_MAX(Y)) / \#(X)) * 100]$

CASE-IV:

Support: $[(\#(X \& Y) / \text{num_transactions}) * 100, ((\#(X \& Y) + P_MAX(X, Y)) / \text{num_transactions}) * 100]$

Confidence: $[(\#(X \& Y) / (\#(X) + P_MAX(X))) * 100, ((\#(X \& Y) + P_MAX(X, Y)) / (\#(X) + P_MAX(X, Y))) * 100]$

Figure 3. Data integrity cases.

significance of establishing intervals of support and confidence is that such information can be used to quantify the integrity of the sanitized data.

3.6 Report Generation

The recorded results include a list of the association rules that violate their security policy and a list of the modified items. We are currently implementing a tool that takes as input an association rule discovered from the sanitized data and outputs a measured range of support and confidence values with respect to the original unsanitized data.

4. Association Mining Risk Assessment Tool (AMRAT)

We have developed a tool, referred to as the Association Mining Risk Assessment Tool (AMRAT), that implements the risk assessment procedure outlined in Section 3. The tool allows a security administrator to systematically alter a collection of data so that there is no violation of a stated security policy. The architecture of AMRAT is shown in Figure 4. The input into the tool is one or more concept hierarchies, a set of association templates and a collection of stored data. The concept hierarchies and association templates are created by the security administrator from the organization's security policies. The output produced by AMRAT includes a collection of sanitized data and a risk assessment report. Currently, AMRAT allows a security administrator to sanitize a collection of data by either concealing items from the data or by altering

Figure 4. Architecture of AMRAT.

item values. The risk assessment report provides a security administrator with a summary of the sanitation process including a list of the "non-secure" rules.

AMRAT is comprised of three components, a concept analyzer, an association analyzer and a data sanitizer. The role of the concept analyzer is to provide the association analyzer with information regarding the structure of the given concept hierarchies. This information includes the individual items that comprise a named concept. AMRAT allows concepts to represent expressions of items defined in terms of the logical operators AND and OR. The role of the association analyzer is to expand the given association templates into individual association rules and to evaluate each of the rules against the collected data. The evaluation process determines if a rule violates its corresponding security policy. AMRAT supports security policies defined in terms of support, confidence and improvement. The results of the association analyzer are utilized by the data sanitizer. The primary role of the sanitizer is to transform the collected data to ensure that there are no violations of security policies. As mentioned above, AMRAT allows a security administrator to choose a particular method for altering the data. The process of altering item values with respect to a specific rule may cause a violation of security policy with respect to other rules. As a result, the sanitizer iteratively evaluates the association rules until all stated rules are evaluated with no reported policy violations.

5. Conclusions and Future Work

In this paper, we proposed a risk assessment procedure for controlling the security and privacy threats presented by association mining. The proposed procedure is based upon a new data analysis process called Knowledge Hiding in Databases (KHD). The goal of KHD is the non-trivial hiding of potentially sensitive knowledge in data. The KHD methodology consists of six steps: identify sensitive knowledge, identify data mining algorithms, formulate security policies, risk assessment, sanitize data, and report generation. These six steps form the basis of our assessment procedure. The results from our initial experiments suggest that the assessment procedure is an effective countermeasure against the security and privacy threats presented by association mining algorithms.

We have several additional research projects planned with respect to this work. Our immediate plans include additional experiments to further evaluate the developed assessment procedure, investigate alternative approaches to sanitizing data, and to extend AMRAT to allow greater control over the risk assessment procedure and to provide more detailed information on the integrity of the sanitized data. We also plan to extend the current work to include the security and privacy threats presented by the discovery of sequential patterns. Finally, we plan to investigate the development of a new data analysis process that is similar to KHD called Knowledge Fabrication in Databases (KFD). The goal of KFD, in contrast to KHD, is to alter a collection of data so that it contains knowledge that is intensionally misleading. A potential application of KFD is the release of data by government agencies in the context of homeland security.

References

- [1] Agrawal, R. and Srikant, R. (2000). Privacy-preserving data mining. *Proceedings of the ACM SIGMOD Conference*.
- [2] Berry, M. and Linoff, G. (1997). *Data Mining Techniques: For Marketing, Sales and Customer Support*, Wiley.
- [3] Chang, L. and Moskowitz, I. (1998). Parsimonious downgrading and decision trees applied to the inference problem. *Proceedings of the New Security Paradigms Workshop*.
- [4] Chang, L. and Moskowitz, I. (2000). An integrated framework for database privacy protection. *Proceedings of the Fourteenth IFIP WG 11.3 Conference on Database and Applications Security*.
- [5] Clifton, C. (1999). Protecting against data mining through samples. *Proceedings of the Thirteenth IFIP WG 11.3 Conference on Database Security*.

- [6] Clifton, C. and Marks, D. (1996). Security and privacy implications of data mining. *Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*.
- [7] Dasseni, E., Verykios, V., Elmagarmid, A. and Bertino, E. (2001). Hiding association rules by using confidence and support. *Proceedings of the Information Hiding Workshop*.
- [8] Johnsten, T. and Raghavan, V. (1999). Impact of decision-region based classification mining algorithms on database security. *Proceedings of the Thirteenth IFIP WG 11.3 Conference on Database Security*.
- [9] Johnsten, T. and Raghavan, V. (2001). Security procedures for classification mining algorithms. *Proceedings of the Fifteenth IFIP WG 11.3 Conference on Database and Applications Security*.
- [10] Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H. and Verkamo, A. (1994). Finding interesting rules from large sets of discovered association rules. *Proceedings of the Third International Conference on Information and Knowledge Management*.
- [11] Leveson, N. (1995). *Safeware: System Safety and Computers*, Addison-Wesley.
- [12] Saysin, Y., Verykios, V. and Clifton, C. (2001) Using unknowns to prevent discovery of association rules. *ACM SIGMOD Record*, vol. 30.