K-anonymous algorithm in protecting privacy in social communication networks

Jiacheng Wang 515030910412

ABSTRACT The rapid development of social communication network has increased the risk in privacy protection, the association between people has become a new weapon of attackers. In the paper I point out that the released dataset of an association rule hiding method may have severe privacy problem since they all achieve to minimize the side effects on the original dataset. An attacker can discover the hidden sensitive association rules with high confidence when there is not enough "blindage". a detailed analysis of the attack is given and I propose a novel association rule hiding metric, K-anonymous. Based on the K-anonymous metric, a framework is presented to hide a group of sensitive association rules while it is guaranteed that the hidden rules are mixed with at least other K-1 rules in the specific region. Several heuristic algorithms are proposed to achieve the hiding process. Experiment results are reported to show the effectiveness and efficiency of the proposed approaches.

Key Words Association Rule Hiding, k-anonymity

1. Introduction

Association rule mining was introduced to discover strong patterns, for example, "people who often communicate on WeChat tend to go out together". Armed with this mining technique, an attacker can make decisions based on how people communicate. Moreover, data sharing can gain mutual benefits to all participants. Data owners usually release their data as Ill as the mining parameters to other partners. However, these advanced technologies have in- creased the risks of disclosing the association rules that the owner considers sensitive when the dataset is shared with other organizations.

To address the problem of preventing the sensitive association rules from being disclosed, researchers have studied methods for Association Rule Hiding. In general, existing approaches sanitize the original dataset such that the sensitive rules cannot be discovered in the released dataset while preserving as much knowledge as possible using the same minimum confidence threshold and minimum support threshold, even if the dataset is shared with other parties.

Example 1: consider that a company wants to distribute its transaction dataset D in Figure 1 to other parties. D has 24 transactions. TID is the index for the transactions. Items is the transaction. The frequent itemsets with support larger than 9 are: DB(10), D(12), HA(10), H(13), IB(10), I(15), A(14), and B(15). The number in the parentheses is the support value for the itemset. t_3 (TI = 3) fully supports AGH and partially supports EG. The support of an itemset X is defined as the numbser of transactions that fully support X, which is denoted as Supp(X). The company uses association rule mining tool to mine the rules using MST (10) and MCT (76.9%). D ⇒ B (Support: 10, Confidence: 83.3%), and H ⇒ A (Support: 10, Confidence: 76.9%) are the two strong rules. The generating set for the rule D ⇒B is DB. The company finds that the rule H ⇒ A is sensitive and wants to hide it. Adopting an existing algorithm, the publisher produces the release dataset D by removing an item "H" in the fourth transaction of D. The rule H ⇒ A is hidden because either its confidence (75%) is less than MCT or its support (9) is less than MST in

dataset D. Using the same MST and MCT, I can only get one rule, that is, D ⇒ B. All existing hiding algorithms try to break the two conditions for an association rule by reducing either the support or the confidence of the sensitive rules.

TID	Items	TID	Items	TID	Items
1	ABCDHI	9	BDEGH	17	ADH
2	BCDFHI	10	AHIJ	18	AHIJ
3	AGHIJ	11	BCDIJ	19	BCDIJ
4	ABDEGHI	12	ABDIJ	20	BCDEGI
5	AHJ	13	ADEFH	21	AB
6	ABCDEHI	14	BDEG	22	BC
7	ABCFI	15	BEGH	23	BEI
8	AGHJ	16	ACEI	24	I

Figure 1

2. Isolation attack

I use a rectangular coordinate system to demonstrate the hiding process in Figure 2. The x-axis represents the support of the association rule while the y-axis represents the confidence of the association rule. A point (s, c) in the system is a rule whose support value is s and whose confidence value is c. The set of association rules from dataset D with MST s and MCT c is denoted as ξ (D, s, c). Any rule in ξ (D, s, c) is called a (s, c)-strong rule with respect to D. Therefore, the (S, C)-strong rules are within the zone Z1.

After applying the association rule hiding algorithms, the sensitive rule r: X ⇒Y, originally in zone Z1, falls into the zone Z2, which is between solid lines and the dotted lines.

Based on the association rule hiding algorithm parameters MST (S) and MCT (C), the adversaries can deduce that the sensitive rules will fall in a certain region. For example, if the adversaries know that the hiding algorithm is to decrease the support of the sensitive rules, and the hiding process needs to minimize the side effect, they can learn that the support for the sensitive rules will be the maximum integer that is less than the given MST. If there is only one rule whose support is equal to the maximum integer in the sanitized dataset, the hidden rule can be identified by the adversaries with 100% confidence. The scenario is like an isolated island in the map which makes it easy to be identified. I call it the isolation attack.

To the best of our knowledge, none of the existing ARH algorithms have addressed this type of attack.

Based on the “minimal impact” principle, I can derive two lower bounds regarding the support value and the confidence value of the sensitive rules after the hiding process.

1). Given MST s and MCT c, the lower bound of the support $s^{\perp}$ for the hidden sensitive rules in D is $s - 1$.

2). Given MST s and MCT c, when adopting confidence based hiding approach, the lower bound of the confidence value $c^{\perp}$ for the hidden sensitive rules is $(c - 1/s)$.

I adopted K-anonymous algorithm, which can be defined as the following: Given the hiding parameter s and c, let $s^{\perp}$ be (s-1) and $c^{\perp}$ be $(c - 1/s)$. The cloak zone M of a sanitized dataset D is the difference between ξ (D, $s^{\perp}$,$c^{\perp}$) and ξ (D, s, c). The cloak zone is exactly the area where the region between the

dotted lines and the solid lines is in Figure 2. I have to point out that there may be other rules rather than the hidden ones in the cloak zone.

An association rule hiding algorithm has K-anonymous property if and only if the number of rules (called size) in the cloak zone M is at least K.

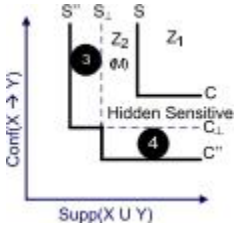


Figure 2

3. Algorithm

I use Figure 2 to intuitively show how my approach, post-sanitization, works. Using existing association rule hiding algorithms, I transform D to D_{hide}, and move the sensitive rules from zone Z1 to zone Z2. If D_{hide} does not satisfy K-anonymous, I obtain the blindage rules from either zone 3 or zone 4 in the figure. The rules in zone 3 is ξ (D, s, c$\perp$) - ξ (D, s$\perp$,c$\perp$), where s is less than s$\perp$. By increasing their support or confidence, the selected rules can move to the cloak zone M (same as Z2) such that the number of rules in M increases. If the sanitized dataset does not satisfy K- anonymity, I promote K blindage rules into the cloak zone instead of making the number of rules in the cloak zone to be K. If I choose to let the number of rules in the cloak zone to be K, I may end up with less than K rules in the zone when some rules fall out of the cloak zone in the sanitization.

I solve the blindage rules problem in three steps. The first step is to define variables xi (i = 1, ..., |S |), which will be 1 if the i-th rule is selected into the result subset, and 0 otherwise. The second one is to build the buckets and place the rules into them. For each distinct item in S, I build a bucket. The set of buckets is denoted as B. For each rule, I put it into the buckets according to the items it supports. I use B_j to denote the j-th bucket. The third step is to derive the constraints and object function as the following:

$$\textbf{maximize}(\sum_{i=1}^{|S|} x_i)$$

$$\text{subject to:} \quad (\sum_{x_j \in B_k} x_j) \leq 1 \qquad \forall B_k \in B, \qquad (1)$$

$$x_i \in \{0,1\} \qquad \forall i \in \{1,...,|S|\}. \qquad (2)$$

The objective function maximizes the number of rules included. Constraint (1) states that no more than one rule can be selected from the same bucket because these rules are overlapped. Constraint (2) imposes the binary requirement on all x_i variables.

After I produce the blindage rules, I have to increase the support (or confidence) value of the blindage rules such that these rules enter the cloak zone. Therefore, the number of rules in the cloak zone increases. This process can be called cloaking. The association rule cloaking algorithms can be described in the Figure 3:

Input: a set R_B of rules to cloak, a dataset D_{hide}, the MCT, the MST, the $MCT_{\perp}$ and the $MST_{\perp}$
Output: the sanitized dataset D'_{hide} from D_{hide} such that the rules in R_B are in the cloak zone of D'_{hide}

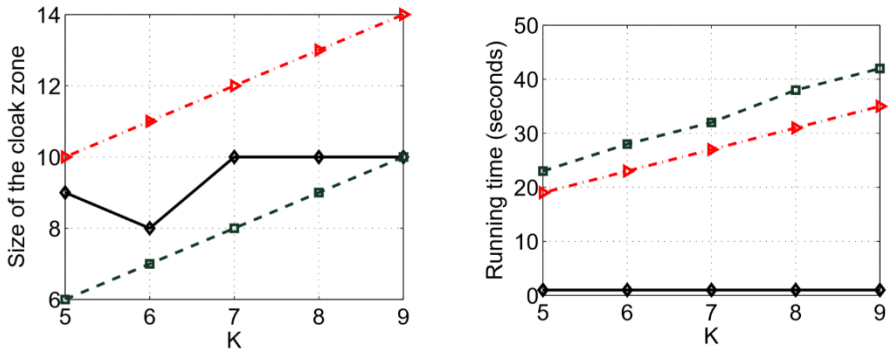
begin
  sort the rules in R_B in descendent order of the support values;
  foreach *rule* $r \in R_B : X \Rightarrow Y$ **do**
    $T_X = \{t \in D$: t fully supports Y and partially supports $X\}$;
    count the number of items in X for every transaction of T_X;
    sort T_X in descending order of the number of items in X;
    Iterations = $MST_{\perp} - Supp(X \cup Y)$;
    for $i = 1$ *to Iterations* **do**
      t = $T_X[0]$;
      add to t all the missing items in X;
      remove t from T_X;
    end
    $R_B = R_B$ - r;
  end
end

Figure 3

4. experiment

I use Enron e-mail dataset as the original dataset.

The results are shown in Figure 4



(a) size of cloak zone (b) running time

Figure 4

The results are in accordance to my theoretical analysis and related researches.

5. Prospect

I think in the future, the K-anonymous algorithm can be widely applied in association-related problems. By using this method, individuals can avoid information leakage themselves in daily social communication. As the technology of mobile internet is updated so fast, certain privacy-protection methods should be emphasized and renewed as well. K-anonymous algorithm can play a larger role in protecting the privacy of users and entrepreneurs in the future.