T-79.514 Special Course on Cryptology

Database randomization via RRT

Isto Niemi
Helsinki University of Technology
isto.niemi@hut.fi

Overview of the Lecture

- Randomized Response
- Association Rule Mining
- Classification of Randomized Data
- Conclusions

Randomized Response

- Developed by Warned in 1965
- The technique tries to solve the problem where responder has attribute A but he dares not share it out to the interviewer.
- Two different models:
  - ⋆ Related-Question Model
  - ⋆ Unrelated-Question Model
- Two questions are asked from the responder instead of one.

Related-Question Model 1/2

- The questions are related so that the answers are opposite to each other.

  1. Do you ever have the sensitive attribute A

  2. Do you never have the sensitive attribute A

- The Responder uses a randomizing device to decide which question to answer without letting the interviewer to know which question is answered.

Related-Question Model 2/2

- The probability of choosing the first question is θ and the second one is $1 - \theta$.
- $P^*(A)$ is the proportion of 'yes'/'no' obtained from the disguised data.
- $P(A)$ is the estimated proportion of the yes'/'no'.
- The estimated proportion can be solved from the equation:

$$P^*(A = yes) = P(A = yes) \cdot \theta + P(A = no) \cdot (1 - \theta)$$
$$P^*(A = no) = P(A = no) \cdot \theta + P(A = yes) \cdot (1 - \theta)$$

Unrelated-Question Model

- Two unrelated questions.
- One of the questions are real.
- The other one is any question with one known probability θ.
- An example:
  1. Flip a coin. Do you ever have the sensitive attribute A?
  2. Flip a coin. Did you get a head?”

Association Rule Mining

- A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In *Proceedings of 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, July 2002.

- They present a framework for mining association rules from the randomized transactions consisting of categorical items.

- They point out the problem of privacy breaches.

- They derive formulae for an unbiased support estimator and its variance.

Uniform Randomization

- Uniform Randomization is presented by Evfimievski et al. as a generalization of Warnes's RRT.
- Before sending a transaction to the server, the responder takes each item and with probability p replaces it by a new one not originally present in this transaction.
- For large values of p, most of the items in transactions are not 'true'. If there are enough transactions frequent itemsets will still be 'visible'.
- Seldom occurred itemsets are problematic because every time seen in a randomized transaction they reveal information of real data.

Privacy Breaches

- Definition: We say that item set A causes a privacy breach of level b if some item $a \in A$ and for some $i \in 1...N$ we have $P[a \in t_i | A \subset t'_i] \geq p$.

- Evfimievski et al. focus on the definition above.

- Ignored information:
  - ⋆ the missing items
  - ⋆ the size of the randomized transactions
  - ⋆ other extra information the server might know

Cut-and-paste Randomization

- An implementation of Uniform Randomization.
- Two input parameters $\rho_m \in (0, 1)$ and $K_m > 0$.
- Handles each transaction t_i independently and converts it to t'_i.
  1. Choose uniformly and random: $j \in (0, K_m)$; if $j > m$ set $j = m$
  2. Select j items out of t_i uniformly at random. These items are placed into t'_i
  3. Each other item (including the rest of t_i) is placed into t'_i with probability ρ_m, independently.

Partial Support

- Let $\mathcal{I} = \{a_1, a_2, ..., a_n\}$, $T = (t_1, t_2, ..., t_N)$ where $t_i \subset \mathcal{I}$ and $A \subseteq \mathcal{I}$
- The fraction of the transactions in T that have intersection with A of size l among all transactions in T is called *partial support* of A for intersection size l:
- $supp_l^T(A) := \frac{\#\{t \in T | \#(A \cap t) = l\}}{N}$
- Support of A is $supp^T(A) = supp_k^T(A)$ for $k = |A|$.

Support Recovery 1/2

- The estimate of the real support is needed for association rule mining.

- For calculations we need P that is the $(k+1) \times (k+1)$ matrix with elements $P_{(l'l)} = p[l \rightarrow l']$.

$$p_k^m[l \rightarrow l'] = p[l \rightarrow l'] := P[\#(t' \cap A) = l' | \#(t \cap A) = l]$$

Support Recovery 2/2

- The expected value of the partial support of the randomized data:

$$\mathbf{E}\vec{s'} = P \cdot \vec{s}$$

- Denote $Q = P^{-1}$ (assume that it exists)

$$\vec{s} = \mathbf{E}Q \cdot \vec{s'}$$

- An unbiased estimator for the original partial supports given randomized partial supports:

$$\vec{s}_{est} := Q \cdot \vec{s'}$$

Limiting Privacy Breaches

- Estimate maximum possible support.
- Given the maximum support, find the situation that are most likely to cause a privacy breach.
- Make randomization just strong enough to prevent such a privacy breach.

Discovering Associations

- The algorithm used is a modified Apriori.
- It is for finding frequent itemsets.
- Shortly (where s_{min} is predefined):
  1. Form all possible k-itemset.
  2. Drop itemset with support belong $s_{min} - \sigma$.
  3. Save itemset for output with support at least s_{min}.
  4. set $k = k + 1$ and go to step 1 with items from not dropped itemsets. Stop if k is too large or no items left.

Experiments of Association Rule Discovery

- Two real life data sets:

  1. The Soccer data set from the clickstream log of the 1998 world cup web site

  2. The mail-order data set consists of items ordered by a customer in a single mail order.

- All long transactions were removed and all too small classes were joined to their parents in the taxonomy.

- Good results: there were only comparatively few false positive and even fewer false drops.

Classification of Randomized Data

- W. Du, and Z. Zhan. Using Randomized Response Techniques for Privacy-Preserving Data Mining. In *Proceedings of 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2003.

- Du et al. present a method to build decision tree classifiers while preserving data's privacy.

Multivariate Randomized Response

- Instead of one question a set of questions is presented to the responder.

- The responder is supposed to either answer all the questions truthfully (with probability θ) or lie to all of them (with probability $1 - \theta$).

$$P^*(E) = P(E) \cdot \theta + P(\overline{E}) \cdot (1 - \theta)$$
$$\overline{P^*(E)} = P(E) \cdot \theta + P(\overline{E}) \cdot (1 - \theta)$$

$$E = (A_1 = 1) \cap (A_2 = 1) \cap (A_3 = 0)$$
$$\overline{E} = (A_1 = 0) \cap (A_2 = 0) \cap (A_3 = 1)$$

Classification

- The aim of classification is to extract models describing important data classes or to predict future data.

- Classification is two-step process:

  1. Model building based on labeled training data

  2. Accuracy calculation

- Decision tree is one of the classification methods.

- It recursively partitions the training data until each partition consists of examples from one class.

Modified ID3 Algorithm

- ID3 is a well known algorithm for decision tree building.
- The difference to the original ID3 algorithm is the Information Gain calculation.
- The attribute A with highest Gain is used to partition the training sample S.
- $Gain(S|A)$ = 'I must transmit S. How many bits on average would it save me if A is known?'

Accuracy Score

- Feed the decision tree with the test record and its' complement.
- If both prediction results are correct then the algorithm is working ok.
- In other situation we can make a conclusion with θ certainties.

$$P^*(correct) = P(correct) \cdot \theta + \overline{P}(correct) \cdot (1 - \theta)$$
$$\overline{P}^*(correct) = \overline{P}(correct) \cdot \theta + P(correct) \cdot (1 - \theta)$$

Classification Experiments

- Prediction task is to determine whether a person's income exceeds $50k/year.
- Data contains 10000 instances with 14 attributes (6 continues and 8 nominal) and it was binarized before usage.
- The experiment shows that with θ in ranges $[0, 0.4]$ abd $[0.6, 1]$ very high accuracy cab still be achieved.

Conclusions: Evfimievski et al.

- Randomization method called Uniform Randomization
- Attributes are independently disguised
- Attributes are randomized before the data is sent to the server
- Best for transactions of categorical data like books bought together
- Maximum size for transactions are around 10
- New algorithm for association rule finding from randomized data

Conclusions: Du et al.

- Randomization method called Multivariate Randomized Response
- Randomization isn't item-invariant
- Attributes are randomized before the data is sent to the server
- No maximum size for attribute vectors
- Aimed for fixed size binary data but can be extended to non binary data like demographic profiles
- New algorithm for classification of randomized data