Result-driven approaches

Vicenç Torra

November, 2022

Umeå University, Sweden

Data Privacy

Respondent and owner privacy

- Data-driven or general-purpose
- Computation-driven or specific-purpose
- Result-driven (Ch. 3.5)

Data Privacy

Result-driven

- Prevent data mining procedures infer some knowledge that is valuable for the database owner
- Other uses: avoid discriminatory knowledge inferred from databases

Data Privacy

Result-driven

- **Formalization.** Database $\mathcal{D}$, A data mining algorithm, with parameters Θ is said to have ability to derive knowledge K from $\mathcal{D}$ if and only if K is obtained from the output of the algorithm. Notation: $(A, \mathcal{D}, \Theta) \vdash K$.
- Any knowledge K such that $(A, \mathcal{D}, \Theta) \vdash K$ is in $KSet_{\mathcal{D}}$.

Data Privacy

Result-driven

- **Formalization.** Database $\mathcal{D}$, A data mining algorithm, with parameters Θ is said to have ability to derive knowledge K from $\mathcal{D}$ if and only if K is obtained from the output of the algorithm. Notation: $(A, \mathcal{D}, \Theta) \vdash K$.
- Any knowledge K such that $(A, \mathcal{D}, \Theta) \vdash K$ is in $KSet_{\mathcal{D}}$.

Definition. $\mathcal{D}$ a database, $\mathcal{K} = \{K_1, \ldots, K_n\}$ sensitive knowledge to be hidden. The problem of hiding knowledge $\mathcal{K}$ from $\mathcal{D}$ consists on transforming $\mathcal{D}$ into a database $\mathcal{D}'$ such that

1. $\mathcal{K} \cap KSet_{\mathcal{D}'} = \emptyset$
2. the information loss from $\mathcal{D}$ to $\mathcal{D}'$ is minimal

Data Privacy

Result-driven for association rules mining: Association rule hiding

- Recall that rules are mined when

$$Support(R) \geq thr - s$$

and

$$Confidence(R) \geq thr - c$$

for certain thresholds $thr - s$ and $thr - c$.

Data Privacy

Result-driven for association rules mining: Association rule hiding

- Recall that rules are mined when

$$Support(R) \geq thr - s$$

and

$$Confidence(R) \geq thr - c$$

for certain thresholds $thr - s$ and $thr - c$.

Two approaches:

- To reduce the support of the rule.
- To reduce the confidence of the rule.

Data Privacy

Result-driven for association rules mining: example

- **A formalization.** $\mathcal{D}$ a database; $thr - s$ threshold. Let $\mathcal{K} = \{K_1, \ldots, K_n\}$ sensitive itemsets, $\mathcal{A}$ non-sensitive itemsets.

Data Privacy

Result-driven for association rules mining: example

- **A formalization.** $\mathcal{D}$ a database; $thr - s$ threshold. Let $\mathcal{K} = \{K_1, \ldots, K_n\}$ sensitive itemsets, $\mathcal{A}$ non-sensitive itemsets.
- Transform $\mathcal{D} \rightarrow \mathcal{D}'$ such that
  1. $Support_{\mathcal{D}'}(K) < thr - s$ for all $K_i \in \mathcal{K}$
  2. The number of itemsets K in $\mathcal{A}$ such that $Support_{\mathcal{D}'}(K) < thr - s$ is minimized.

This problem is NP-hard (Atallah et al., 1999)

Because of this: heuristic approaches

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Algorithm.**

  While HI is not hidden **do**
  HI' = HI;
  While $|HI'| > 2$ **do**
  P = subsets of HI with cardinality $|HI'| - 1$;
  HI'= $\arg\max_{hi \in P} Support(hi)$;
  Ts = transaction in T supporting HI that affects
  the mininum number of itemsets of cardinality 2;
  Set HI' = 0 in Ts;
  Propagate results forward;

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Algorithm.**

  While HI is not hidden **do**
  HI' = HI;
  While $|HI'| > 2$ **do**
  P = subsets of HI with cardinality $|HI'| - 1$;
  HI'= $\arg\max_{hi \in P} Support(hi)$;
  Ts = transaction in T supporting HI that affects the mininum number of itemsets of cardinality 2;
  Set HI' = 0 in Ts;
  Propagate results forward;

- The algorithm does not cause false positives,
- only false negatives (rules no longer inferred)

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.

Transaction number	Items
T1	a, b, c, d
T2	a, b, c
T3	a, c, d

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.

Transaction number	Items
T1	a, b, c, d
T2	a, b, c
T3	a, c, d

  - Subsets of HI with cardinality $|HI| - 1$: $\{a, b\}$, $\{b, c\}$, $\{a, c\}$.

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.

Transaction number	Items
T1	a, b, c, d
T2	a, b, c
T3	a, c, d

  - Subsets of HI with cardinality $|HI| - 1$: $\{a, b\}$, $\{b, c\}$, $\{a, c\}$.
  - $Support(\{a, b\}) = Support(\{b, c\}) = 2$, and $Support(\{a, c\}) = 3$

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.

Transaction number	Items
T1	a, b, c, d
T2	a, b, c
T3	a, c, d

  - Subsets of HI with cardinality $|HI| - 1$: $\{a, b\}$, $\{b, c\}$, $\{a, c\}$.
  - $Support(\{a, b\}) = Support(\{b, c\}) = 2$, and $Support(\{a, c\}) = 3$
    $\rightarrow$ We select $HI' = \{a, c\}$.

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.

Transaction number	Items
T1	a, b, c, d
T2	a, b, c
T3	a, c, d

  - Subsets of HI with cardinality $|HI| - 1$: $\{a, b\}$, $\{b, c\}$, $\{a, c\}$.
  - $Support(\{a, b\}) = Support(\{b, c\}) = 2$, and $Support(\{a, c\}) = 3$ $\rightarrow$ We select $HI' = \{a, c\}$.
  - Set of transactions in T that support HI (and HI'): $\{T1, T2\}$.

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.

Transaction number	Items
T1	a, b, c, d
T2	a, b, c
T3	a, c, d

  - Subsets of HI with cardinality $|HI| - 1$: $\{a, b\}$, $\{b, c\}$, $\{a, c\}$.
  - $Support(\{a, b\}) = Support(\{b, c\}) = 2$, and $Support(\{a, c\}) = 3$ $\rightarrow$ We select $HI' = \{a, c\}$.
  - Set of transactions in T that support HI (and HI'): $\{T1, T2\}$.
  - Ts transaction in $\{T1, T2\}$ that affects the minimum number of itemsets of cardinality 2: $T2$ affects less itemsets than $T1$.

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.
- Remove one of the items in $HI' = \{a, c\}$ that are in $T2$:

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.
- Remove one of the items in $HI' = \{a, c\}$ that are in $T2$:
  Both have the same support, we select one of them at random.
- Propagate the results forward: recompute supports