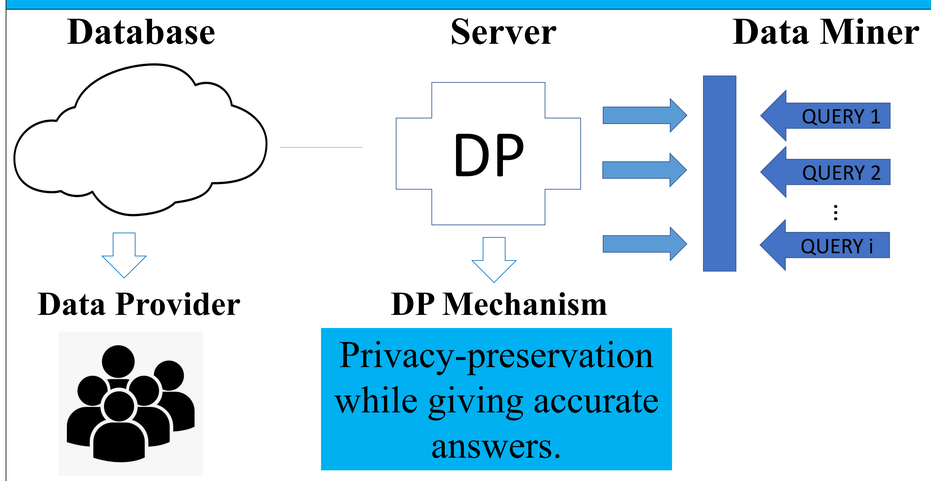
DIFFERENTIALLY PRIVATE GREEDY DECISION FOREST

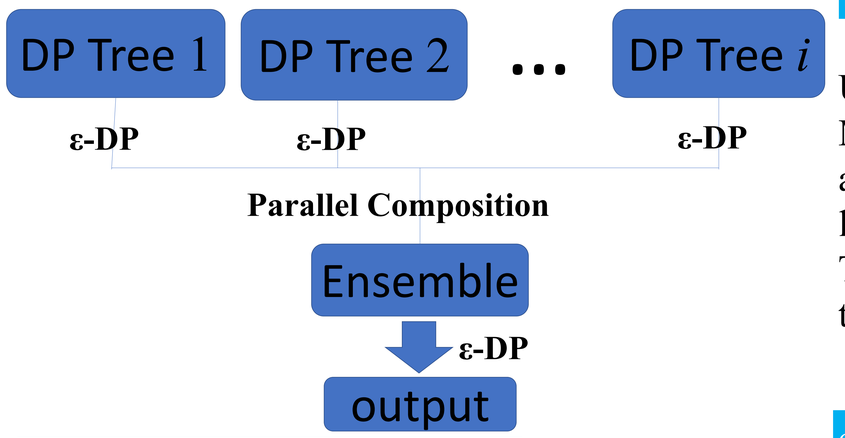
Bangzhou Xin, Wei Yang, Shaowei Wang, Liusheng Huang
University of Science and Technology of China

1. Motivation



Privacy-preservation while giving accurate answers.

2. Model structure



a) Parameter selection

- Tree depth: d = [*m*/2, *m*], *m* means the number of attributes in the dataset
- Tree number: Dynamically adjust the number based on the dataset

b) Create a single decision tree

Using Gini index with Exponential Mechanism to choose splitting attributes, and using the smooth sensitivity at the leaf nodes to reduce the influence of noise. The greedy differentially private decision tree is finally obtained.

c) Distribution of privacy budget

- ✓ Each tree gets the whole ε privacy budget for the parallel composition
- ✓ Layer privacy budget = $\frac{\varepsilon}{2\times(d-1)}$
- ✓ Leaf layer privacy budget = $\frac{\varepsilon}{2}$

3. Theoretical analysis

Theorem . Differentially Private Greedy Decision Forest satisfies ε-differential privacy.

For each layer, it satisfies DP to choose splitting attributes with probability:

$$\Pr(a) \propto \exp\left(\frac{-\varepsilon \times Gini(a,A)}{2(d-1)\times\Delta(Gini)}\right)$$

and all attribute selections follow the Composition Theory. All internal layers satisfy α−differential privacy, where

$$\alpha = \sum_{i=1}^{d-1}\varepsilon_i = (d-1)\times\frac{\varepsilon}{2\times(d-1)} = \frac{\varepsilon}{2}$$

Combining with leaf layer, each tree satisfies ε -differential privacy. Because the datasets used in each tree are not adjacent, the whole forest is ε -differential privacy.

4. Experiment

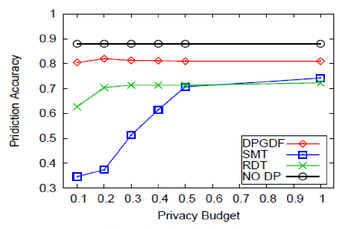


Figure 1 : Car Evaluation

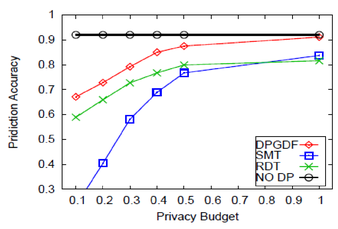


Figure 3 : Nursery

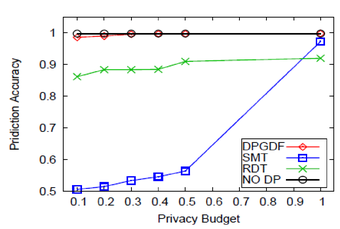


Figure 2 : Mushroom

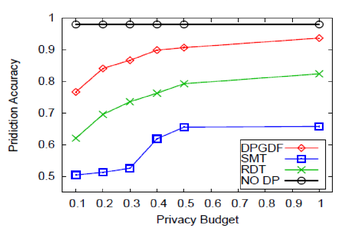


Figure 4 : Chess

5. Conclusions and Contribution

Conclusion: Testing with public datasets in UCI Machine Learning Repository, the accuracy of our algorithm is always 10 percentage points higher than other algorithms under the same privacy budget.

Contribution: DPGDF can be applied to the datasets query which containing privacy information. While offering privacy-preservation, it can also give more accurate query results.