Data Anonymization that Leads to the Most Accurate Estimates of Statistical Characteristics

Gang Xiang[1] and Vladik Kreinovich[2]

[1]Applied Biomathematics, 100 North Country Rd.
Setauket, NY 11733, USA, gxiang@sigmaxi.net

[2]Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968, USA, vladik@utep.edu

1. Need to Preserve Privacy

- One of the main objectives of engineering is to help people:
  - civil engineering designs houses in which we live and roads along which we travel,
  - electrical engineering designs appliances – and electric networks that help use these appliances.
- To better serve customers, it is important to know as much as possible about the potential customers.
- Customers are reluctant to share information, since this information can be potentially used against them.
- For example, age can be used by companies to (unlawfully) discriminate against older job applicants.
- It is thus important to preserve privacy when storing customer data.

2. How to Preserve Privacy: k-Anonymity and ℓ-Diversity

- To maintain privacy, we divide the space of all possible combinations of values $(x_1, \ldots, x_n)$ into boxes.
- For each record, instead of storing the actual values x_i, we only store the label of the box containing x.
- To avoid further loss of privacy, it is important to make sure that location in a box does not identify a person.
- This is usually achieved by requiring that for some fixed k, each box contains at least k records.
- It is also not good if all records within a box have the same value of an i-th quantity x_i.
- It is thus required that for some ℓ, in each box there are at least ℓ different values of each x_i.

3. Statistical Data Processing

- Our main objective is to predict the desired characteristic x_{i_0}.
- In most cases, the dependence is linear, so we must find c_q s.t. $x_{i_0} \approx c_0 + \sum_{q=1}^{m} c_q \cdot x_{i_q}$.
- Least Squares Approach leads to:

$$\sum_{r=1}^{m} c_r \cdot C_{i_q i_r} = C_{i_0 i_q}; \quad c_0 = E_{i_0} - \sum_{q=1}^{N} c_q \cdot E_{i_q}.$$

- We also want to know which quantities are correlated, i.e., we want to estimate $\rho_{ij} = \dfrac{C_{ij}}{\sigma_i \cdot \sigma_j}$.
- In all these tasks, we need to estimate averages E_i, variances $V_i = \sigma_i^2$, covariances C_{ij}, and correlations ρ_{ij}.

4. Statistical Characteristics: Reminder

- The means are usually estimated as follows:

$$E_i = \frac{1}{N} \cdot \sum_{p=1}^{N} x_i^{(p)}, \quad E_j = \frac{1}{N} \cdot \sum_{p=1}^{N} x_j^{(p)}.$$

- The covariance is usually estimated as:

$$C_{ij} = \frac{1}{N} \cdot \sum_{p=1}^{N} \left(x_i^{(p)} - E_i \right) \cdot \left(x_j^{(p)} - E_j \right).$$

- The variance is usually estimated as:

$$V_i = \frac{1}{N} \cdot \sum_{p=1}^{N} \left(x_i^{(p)} - E_i \right)^2.$$

5. In Statistical Data Processing, Privacy Leads to Uncertainty

- To maintain privacy, we replace each numerical value $x_i^{(p)}$ with the corresponding interval.
- Different values from these intervals lead to different values of the resulting statistical characteristics.
- Hence, for each characteristic, we get a whole interval of possible values.
- If this interval is too wide, the resulting range is useless: e.g., $[-1, 1]$ for correlation.
- It is therefore desirable to select:
  - among all possible subdivisions into boxes which preserve k-anonymity (and ℓ-diversity),
  - the one which leads to the narrowest intervals for the desired statistical characteristic.

6. Estimating Accuracy Caused by Privacy-Based Subdivision into Boxes: Case of k-Anonymity

- To minimize uncertainty, we select the smallest boxes.
- Hence, each box B should have exactly k records.
- For intervals $[\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$, instead of $C(x_1^{(1)}, \ldots, x_n^{(N)})$, we get:

  $C(\widetilde{x}_1^{(1)} + \Delta x_1^{(1)}, \ldots, \widetilde{x}_n^{(N)} + \Delta x_n^{(N)})$, where $|\Delta x_i^{(p)}| \le \Delta_i$.
- When we have many records, boxes are small, so we can use a linear approximation:

$$C = \widetilde{C} + \sum_{p=1}^{N} \sum_{i=1}^{n} \frac{\partial C}{\partial x_i} \cdot \Delta x_i^{(p)}.$$

- The range of this linear expression is $[\widetilde{C} - \Delta, \widetilde{C} + \Delta]$, where $\Delta \stackrel{\text{def}}{=} \sum_{p=1}^{N} \sum_{i=1}^{n} \left| \frac{\partial C}{\partial x_i} \right| \cdot \Delta_i^{(p)} = k \cdot \sum_{B} \sum_{i=1}^{n} \left| \frac{\partial C}{\partial x_i} \right| \cdot \Delta_i$.

7. Expressions for the Corr. Partial Derivatives

- The estimate for the accuracy Δ is described in terms of partial derivatives $\dfrac{\partial C}{\partial x_i}$ of the stat. characteristic C.
- For the mean E_i, the derivative is equal to $\dfrac{\partial E_i}{\partial x_i} = \dfrac{1}{N}$.
- For the variance V_i, we have $\dfrac{\partial V_i}{\partial x_i} = \dfrac{2 \cdot (x_i - E_i)}{N}$.
- Therefore, for $\sigma_i = \sqrt{V_i}$, we get $\dfrac{\partial \sigma_i}{\partial x_i} = \dfrac{x_i - E_x}{\sigma_x}$.
- For the covariance C_{ij}, we have $\dfrac{\partial C_{ij}}{\partial x_i} = \dfrac{x_j - E_j}{N}$.
- For the correlation ρ_{ij}, we have:

$$\frac{\partial \rho_{ij}}{\partial x_i} = \frac{1}{N} \cdot \frac{(x_j - E_j) - \dfrac{C_{ij}}{\sigma_i^2} \cdot (x_i - E_i)}{\sigma_i \cdot \sigma_j}.$$

8. Towards Optimal Subdivision into Boxes

- The overall expression for Δ is a sum of terms corresponding to different points.
- So, to minimize Δ, we must, for each point, minimize the corresponding term $\sum_{i=1}^{n} a_i \cdot \Delta_i$, where $a_i \stackrel{\text{def}}{=} \left|\frac{\partial C}{\partial x_i}\right|$.
- The only constraint on the values Δ_i is that the corresponding box should contain exactly k different points.
- The number of points can be obtained by multiplying the data density $\rho(x)$ by the box volume $\prod_{i=1}^{n}(2\Delta_i)$.
- The data density can be estimated based on the data.
- So, we minimize $\sum_{i=1}^{n} a_i \cdot \Delta_i$ under the constraint

$$\rho(x) \cdot 2^n \cdot \prod_{i=1}^{n} \Delta_i = k.$$

9. First Result: (Asymptotically) Optimal Subdivision into Boxes (Case of k-Anonymity)

- *Method:* Lagrange multiplier technique leads to

$$\Delta_i = \frac{c(x)}{a_i}, \text{ where } a_i = \left|\frac{\partial C}{\partial x_i}\right|.$$

- From the constraint, we get $c(x) = \frac{1}{2} \cdot \sqrt[n]{\frac{k}{\rho(x)} \cdot \prod_{j=1}^{n} a_j}$.

- *Conclusion:* around each point x, we need to select the box with half-widths

$$\Delta_i = \frac{1}{2} \cdot \sqrt[n]{\frac{k}{\rho(x)}} \cdot \frac{\sqrt[n]{\prod_{j=1}^{n} a_j}}{a_i}.$$

- *The resulting accuracy:* $\Delta = n \cdot \sum_x c(x)$, where the sum is taken over all N data points x.

10. We Need to Dismiss Rare Points

- In many practical situations, we have rare points, for which the smallest box containing k of them is huge.
- This big-size box will contribute a large amount of uncertainty to Δ; so we should dismiss such rare points.
- If we select a subset $S \subset \{1, 2, \ldots, N\}$ of the set of N original points, then:
  - the privacy-related uncertainty reduces to $n \cdot \sum_{x \in S} c(x)$,
  - but the stat. accuracy reduces to $\dfrac{A}{\#(S)}$.
- Minimizing $n \cdot \sum_{x \in S} c(x) + \dfrac{A}{\sqrt{\#(S)}}$ leads to selecting all x with $c(x) \le c_0$, where c_0 minimizes the sum

$$n \cdot \sum_{x : c(x) \le c_0} c(x) + \frac{A}{\sqrt{\#\{x : c(x) \le c_0\}}}.$$

11. Examples

- For estimating the mean E_i, we have $a_i = \text{const}$ and thus, $c(x) = \text{const} \cdot \dfrac{1}{\sqrt[n]{\rho(x)}}$.
- In this case, $c(x)$ is a decreasing function of density.
- So dismissing points with $c(x) > c_0$ is equivalent to dismissing all the points with $\rho(x) < \rho_0$ (for some ρ_0).
- For computing covariance C_{ij}, the derivative a_i is proportional to $x_i - E_i$.
- So, the upper threshold c_0 on $c(x)$ is equivalent to the lower threshold on the ratio $\dfrac{\rho(x)}{|x_i - E_i| \cdot |x_j - E_j|}$.
- Thus, we can also use points x with small $\rho(x)$ – if x_i or x_j is close to the corresponding mean.
- Using extra points x improves accuracy.

12. How to Also Take into Account ℓ-Diversity

- Up to now, we only took into account the k-anonymity requirement.
- We also need to take into account that within each box, for each variable x_i, there are $\geq \ell$ different values of x_i.
- To formalize this requirement, we first need to describe what "different" means.
- Usually, for each variable i, different means that $|x_i - x_i'| \geq \varepsilon_i$ for some threshold ε_i.
- Thus, ℓ different values means that $2\Delta_i \geq \ell \cdot \varepsilon_i$.
- *Problem:* find Δ_i s.t. $\sum_{i=1}^{n} a_i \cdot \Delta_i \to \min$ under the constraints $\prod_{i=1}^{n} \Delta_i \geq \dfrac{k}{2^n \cdot \rho(x)}$ and $2\Delta_i \geq \ell \cdot \varepsilon_i$ for all i.

13. Main Result: Optimal Subdivision into Boxes

- Around each point x, we first compute the values

$$\Delta_i = \frac{1}{2} \cdot \sqrt[n]{\frac{k}{\rho(x)}} \cdot \frac{\sqrt[n]{\prod_{j=1}^{n} a_j}}{a_i}, \text{ where } a_i = \left| \frac{\partial C}{\partial x_i} \right|.$$

- If $2\Delta_i \geq \ell \cdot \varepsilon_i$ for all i, we select Δ_i.
- Otherwise, we sort the quantities by $a_i \cdot \varepsilon_i$:

$$a_1 \cdot \varepsilon_1 \geq a_2 \cdot \varepsilon_2 \geq \ldots \geq a_n \cdot \varepsilon_n.$$

- Then, for each t from 1 to n, we compute

$$c_t = \frac{1}{2} \cdot \left(\frac{k \cdot \prod_{i=t+1}^{n} a_i}{\rho(x) \cdot \ell^t \cdot \prod_{i=1}^{t} \varepsilon_i} \right)^{1/(n-t)}.$$

14. Main Result (cont-d)

- For each t, if $\dfrac{2c_t}{\ell} \geq a_{t+1} \cdot \varepsilon_{t+1}$, we compute

$$\Delta(t) \stackrel{\text{def}}{=} \frac{1}{2} \cdot \ell \cdot \sum_{i=1}^{t} a_i \cdot \varepsilon_i + (n-t) \cdot c_t.$$

- We select t s.t. $\Delta(t) \to \min$, and take $\Delta_i = \dfrac{1}{2} \cdot \ell \cdot \varepsilon_i$ for $i \leq t$, and $\Delta_i = \dfrac{c_t}{a_i}$ for $i > t$.
- *Comment.* The computation time of this algorithm is quadratic in n.
- This is OK, since the number n of different characteristics is usually reasonably small.
- What is important is that the algorithm is still linear-time in terms of the number of records N.

15. From an Asymptotically Optimal Anonymization to an Optimal One

- Often, in practice, we have a huge amount of data.
- In such cases, the corresponding boxes containing k records are small.
- In this case, the approximate expression for uncertainty is almost equal to the exact one.
- So, when we minimize the approximate expression, we thus, in effect, minimize the actual uncertainty as well.
- However, in many practical situations, the amount of data is not as huge and thus, boxes are not as small.
- In such situations, our asymptotically optimal partition provides only an approximate optimum.
- In such situations, it is desirable to try to find the actual optimum.

16. Need for Computational Intelligence Techniques

- When we only took linear terms into account, we were able to get an almost explicit analytical solution.
- Once we take quadratic terms into account, the optimization problem becomes NP-hard.
- In practice, we can solve *some* NP-hard problems – if we use additional expert knowledge.
- In other words, we need to use *computational intelligence* techniques.
- Thus, to get from asymptotically optimal to optimal partitions, we need to use computational intelligence.

17. Which Computational Intelligence Techniques Can We Use?

- The three main classes of computational intelligence techniques are:
  - fuzzy logic techniques, that enable us to formalize imprecise ("fuzzy") expert knowledge;
  - neural network techniques, that enable us to learn new techniques and new ideas;
  - techniques of evolutionary computation which enable us to optimize.
- Since our main objective is optimization, a natural idea is to use *evolutionary computation* techniques.
- Also, to capture expert knowledge, it is reasonable to use fuzzy techniques.
- In our future work, we plan to use computational intelligence techniques.

18. Acknowledgments

- This work was supported in part:
  - by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721,
  - by Grant 1 T36 GM078000-01 and grant "Balancing disclosure risk with inferential power: software for intervalized data" from the National Institutes of Health, and
  - by a grant on F-transforms from the Office of Naval Research.
- The authors are thankful to Scott Ferson, Lev Ginzburg, and Luc Longpré for valuable discussions.

19. Bibliography on Anonymization

- G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "A Framework for Efficient Data Anonymization under Privacy and Accuracy Constraints", *ACM Transactions on Database Systems*, 2009, Vol. 34, No. 2, Article 9.
- L. Sweeney, "k-anonymity: a model for protecting privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-Based System*, 2002, Vol. 10, No. 5, pp. 557–570.

20. Bibliography on Statistics and Optimization

- V. Kreinovich, A. Lakeyev, J. Rohn, and P. Kahl, *Computational Complexity and Feasibility of Data Processing and Interval Computations*, Kluwer, Dordrecht, 1997.
- H. T. Nguyen, V. Kreinovich, B. Wu, and G. Xiang, *Computing Statistics under Interval and Fuzzy Uncertainty*, Springer Verlag, 2012.
- P. M. Pardalos, *Complexity in Numerical Optimization*, World Scientific, Singapore, 1993.
- D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, Boca Raton, Florida, 2007.

21. Bibliography on Computational Intelligence

- A. E. Eiben and J. E. Smith, *Introduction to Evolutionary Computing*, Springer Verlag, Berlin, Heidelberg, 2010.
- A. P. Engelbrecht, *Computational Intelligence: An Introduction*, Wiley, Chichester, England, UK, 2007.
- G. J. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic*, Prentice Hall, Upper Saddle River, New Jersey, 1995.
- H. T. Nguyen and E. A. Walker, *First Course In Fuzzy Logic*, CRC Press, Boca Raton, Florida, 2006.
- L. Rutkowski, *Computational Intelligence: Methods and Techniques*, Springer Verlag, Berlin, Heidelberg, 2010.
- L. A. Zadeh, "Fuzzy sets", *Information and control*, 1965, Vol. 8, pp. 338–353.