DaSe Lab
Data Semantics Laboratory
Dept. of Computer Science and Engineering

Data Perturbation via Randomized Normalization for Privacy Protection

Reihaneh Amini, Michelle Cheatham and Nazifa Karima
DaSe Lab, Wright State University, OH, U.S.A

Introduction

Significant research on data perturbation has been done; however, the critical point is that it is not easy to avoid losing accuracy while trying to maximize privacy.

We discuss a way to improve the previous work [1], in which the authors attempt to balance between privacy and utility through the use of Min Max Normalization together with a shifting factor to perturb the data. This technique does a good job of preserving data utility but suffers from a major weakness (lack of randomness) that endangers privacy.

Our technique perturbs and permutes numerical data points while still maintaining the accuracy of data mining techniques after distortion.

[1] Aggarwal, Charu C., and S. Yu Philip. *A general survey of privacy-preserving data mining models and algorithms*. Springer US, 2008..

Approach and Uniqueness

Both the original method and our proposed modification begin with a 2D matrix D containing the original dataset. Matrix D is normalized by Min Max normalization to produce another matrix called $\bar{D}$. $\bar{D}$ is the normalized matrix with the same number of columns and rows, but the values of the data points in $\bar{D}$ are all between 0 and 1.

Original Approach:

The next step in the original approach is to multiply each value in $\bar{D}$ by a constant negative value. This shifts the normalized data into a new negative range and reflects it across the horizontal axis.

Our Approach:

In contrast, our approach does not multiply each value in the normalized matrix by a single constant, but instead by a distinct random number, RN, chosen from a uniform distribution in the range [0.95, 1.05]. The mean of this random noise interval is close to one, which preserves data utility for most applications. Also, the width of the interval can be chosen by the data publisher.

Results and Discussion:

Figure 1 shows the distribution of the perturbed data according to the method described in [1] and our proposed method.

Figure 2 shows the result of an attack based on knowledge of the minimum and maximum values in the dataset, as well as of a single data point. Because the original approach applies the same perturbation process has been applied to every point in the dataset, the attacker is able to completely undo the process and recover the exact original values for every point. In essence, the **lack of randomness** inherent in this approach makes it very brittle.

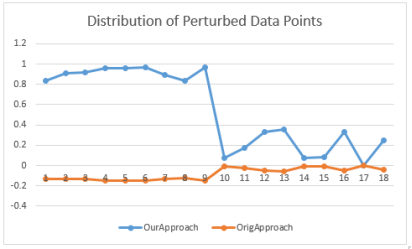

Figure 1. Distribution of Perturbed Data Points by the Original and Proposed Methods.

We contend that every technique used when perturbing datasets must have some degree of randomness in order to be useful in preventing information from leaking to attackers. We follow this guiding principle in our approach to perturbation.

Figure 2 also shows the result of the same attack on the dataset when perturbed according to our proposed method. For this demonstration we have used a uniform random distribution with a range of [0.95, 1.05]. It is evident that the recovered data values are different from the original data, while the overall distribution is similar.

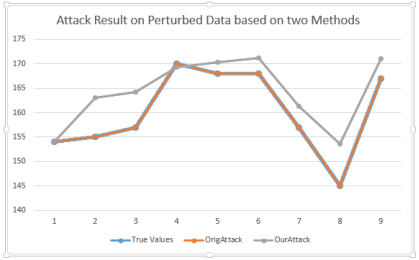

Figure 2. Distribution of Recovered Data Points by the Original and Proposed Methods.

In general, our approach has the following benefits:

- Because the **mean** of the generated random numbers is close to one and their range is limited, the utility of the data is maintained.
- Knowledge of the real value corresponding to a single perturbed data point is no longer sufficient to recover all of original values exactly.
- Adjusting the width of the interval allows the data publisher to control the degree of perturbation, allowing them to choose the appropriate **tradeoff between accuracy and privacy** for their particular datasets: a narrow interval will perturb the data less than a wider one.

Conclusion

A data perturbation algorithm that relies on one or two constant parameters is not secure. To effectively preserve data privacy we should add some randomness to our data; however, we need to be careful about the distribution of this random noise, because it has a direct effect on data mining utility and accuracy. Our proposed method is appropriate for perturbing numerical data, particularly if the accuracy of data mining is of concern. The major limitation of our method is that it cannot deal with binary and categorical datasets; thus this can be the target of our future research work.