Deriving Private Information from Randomized Data - Review

Shumin Guo

Literature papers on privacy preserving data mining have proved that randomization can be used to preserve privacy, but there are critics that randomization methods may not be sufficient to preserve privacy. So this paper is trying to study reasons that lead to the weakness of privacy preservation of randomization methods, and to understand how these features affect the privacy preservation property. It also discussed properties of data and randomization that will have higher risk of disclosing privacy content even though they are randomized.

In this paper, the author assert that a variety of information can lead to the disclosure of private information including attribute dependency, sample dependency, partial value disclosure and the results of data mining operations. This paper focuses on the correlations among attributes of data set. And it recommends three corelation based methods to prove this proposal.

Univariate data reconstruction scheme is a kind of algorithm which does not utilize the correlations among data attributes. It is proposed as a baseline method in this paper. As it only considers the distribution of only one dimension, so if the attributes are highly correlated, the risk of breaking privacy will be high.

The second method is based on Principle Component Analysis (PCA), which allows us to control the correlation among attributes, so that we can study the relationships between correlation and privacy.

Bayes estimate is proposed as a more general solution to this problem. With Bayes estimate, original data is estimated from the randomized data according to the rule that posterior probability be maximized. Experimental results show that this method can achieve better performance than the PCA-based method.

In order to study the properties of data and randomization that lead to higher risk of privacy, a modified randomization scheme is proposed to control the similarity between the original data and randomization noise so that we can study the privacy preservation with different similarity between these two.

Experimental results show that higher correlation among data set attributes can make privacy preservation less effective. Thus proved the assert that correlation have effect on the privacy level.

And results also show that the higher the similarity between data and noise, the less accuracy for data reconstruction, and thus the higher privacy preservation. Based on this experimental result, an improved randomization scheme is proposed to enforce privacy preservation. This scheme is trying to make the original data and the noise have similar covariance matrix, and make them both concentrate on the principal component of the original data so that higher privacy level can be achieved.

This paper is trying to analyze the factors that will make privacy preservation less effective. Several factors are proposed and the correlation among data set attributes are focused. In order to prove and quantify the assertion of the efficacy of correlation among data attributes, methematical proof are given. Extensive experimental results show that the correlation of data attributes can cause the leakage of private information.

The defect of this paper is that computation algorithms are not given and the computational performance and complexity are not discussed.