Abstract

Privacy-preserving in data mining refers to the area of data mining that seeks to safeguard sensitive information from unsolicited or unsanctioned disclosure and hence protecting individual data records and their privacy. Data perturbation is a privacy preservation technique which does addition / multiplication of noise to the original data. It performs anonymization based on the data type of sensitive data. Generalization is a technique were quasi identifiers data are replaced by some other more general term. In this paper privacy protection is applied to high dimensional datasets like Adult and Census. For ranking the attributes, information gain feature subset selection method is used. The high ranking attributes with sensitive information are set as quasi identifiers of the datasets. A hybrid perturbation technique is used to perturb categorical and numeric attributes of both the datasets and the utility of the datasets is measured using accuracy on data mining functionalities. The data distortion is measured using maintenance of Rank of Features (CK) between the original and perturb datasets. Experimental results show that utility of the perturbed datasets comparable with the original dataset and the Census dataset has comparable CK value than adult dataset.
References

- Alexandre Evfimievski : Privacy-Preserving Data Mining by IBM Almaden Research Center, USA Tyrone Grandison IBM Almaden Research Center, USA.
- Mark Hall.; Eibe Frank,; Geoffrey Holmes,; Bernhard Pfahringer,; Peter Reutemann,; Ian H. Witten (2009): The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1
- Pengpeng Lin.; Jun Zhang,; Ingrid St. Omer,; Huanjing Wang.; JieWang Proceedings(2011: A Comparative study on Data perturbation with feature selection, The international multi conference of Engineers and computer scientist 2011 vol 1, March 16-18,
Hybrid Perturbation Technique using Feature Selection Method for Privacy Preservation in Data Mining

2011 Hong Kong.

**Index Terms**

Computer Science  
Information Systems

**Keywords**

Data mining  Privacy preservation perturbation generalization utility classifications

clustering

maintenance of Rank of Features