FEATURE SELECTION IN PRIVACY PRESERVING IN DATA MINING

By:
Kavisha Patel
Enrollment No.: 140370702015

Guided by:
Mr. Amit Rathod
M.E (IT), Assistant Professor

A **Thesis** Submitted to
Gujarat Technological University in Partial Fulfillment of the Requirements for the Master of Engineering in **Computer Engineering**

May – 2016

Computer Science & Engineering Department,
Parul Institute of Engineering & Technology
P.O: Limda, Ta. Waghodia, Dist.: Vadodara

Feature Selection in Privacy Preserving in Data Mining

Submitted By

Kavisha Patel

Supervised By

Mr. Amit Rathod

M.E. (IT), Assistant Professor

Parul Institute of Engineering and

Technology, Limda, Vaghodia, Vadodara

ABSTRACT

Data mining is the pulling out of veiled data from large database. The main problem that rises in any bulk group of data is that of secrecy of the data. Privacy-preserving in data mining (PPDM) is the part of data mining that pursues to protect sensitive information from unwanted or illegal disclosure. The attributes are segregated based on their sensitivity for privacy preservation purposes. Automating this attribute segregation becomes complicated in high dimensional datasets and data streams. Feature selection is an important aspect in privacy preserving. In the Existing System, information or correlation of the attribute on the target class attribute is measured using Information Gain [IG], Gain Ratio [GR] and Pearson Correlation [PC] ranker based feature selection methods with decision tree and this values are used to segregate them as Sensitive Attributes [SA], Quasi Identifiers [QI] and Non-Sensitive. Segregated attributes are subjected to various levels of privacy preservation using both the Double layer Perturbation [DLP] and Single Layer Perturbation [SLP] algorithms to form the perturbed datasets. Since the attribute selection uses tree structure in methods like IG and GR for each attribute value selection in multiple passes, its time complexity is higher. The dissertation work proposes to use the linked array for calculating information gain that would involve only one-pass reading of the dataset and hence increase the efficiency.