ABSTRACT

Corporations are increasingly maintaining large size electronic databases, which are accessed using internet or intranet. Important information utilized from the data using data mining methods. While performing data mining steps, there is an inherent danger to the privacy of the data. The sensitive data available in the database should not be accessible to all users. Techniques proposed in the literature for privacy protection. Privacy-Preserving Data Mining (PPDM) use algorithms on confidential data that are unknown even to the algorithm operator. Users’ personal information and information concerning their collective activity are two major privacy preservation dimensions. The former is termed individual privacy preservation and the latter collective privacy preservation.

Most of the privacy preserving methods are based on reduction in the granularity of the representation of the data. This leads to information loss but improves privacy. Therefore, in PPDM there is a trade-off between loss of information and the privacy. Effective techniques are required which do not compromise the security mechanisms. Some of the techniques proposed for privacy preserving include randomization method, k-anonymity model, l-diversity and distributed privacy preservation. The k-anonymity model is based on a quasi-identifier, which is a collection of attributes in a database that is the identifier for the entire data. All the data in the database is assumed
to be in a set of tables, and each tuple is information of an individual customer.

K-anonymity techniques are based on the reduction of granularity in representation of data using pseudo-identifiers. Major techniques used for granularity reduction are generalization and suppression. In generalization, the attribute values are converted into a range that reduces the granularity and reduces the risk of identifying individual values. In suppression method, actual value of the attribute is removed completely. But these two methods introduce loss of some detail which may affect the accuracy. This motivates the search for anonymization algorithms that achieve the required level of anonymization while incurring a minimal loss of information.

Finding optimal k-anonymous datasets using generalization or suppression has been proved to be a NP-hard problem. Therefore, some standard heuristic search techniques such as Genetic Algorithms (GAs), Particle Swam Optimization (PSO) and Ant Colony Optimization (ACO) can be used to find optimal datasets. Following are the objectives of the research:

- The effect of the anonymization due to k-anonymity on the data mining classifiers is investigated.
- Optimize search for right tradeoff between privacy and information loss using Genetic Algorithm (GA) is proposed for anonymization.
- Hybrid optimization based on Simulated Annealing (SA) with GA is proposed to preserve the classification accuracy.
In the initial stages of research, the anonymization effect due to k-anonymity on the data mining classifiers was investigated. Data is anonymized for different granularity. Mushroom dataset and IPUMS dataset were used for evaluation. Naïve Bayes classifier evaluated anonymized and non-anonymized data with results showing that anonymity increase lead to proportional degradation of classifier performance. The classification accuracy of the Naïve Bayes classifier reduces in the range of 0.41% to 5.33% as the anonymity level increases from 5 to 50 for Mushroom dataset and 0.23% to 4.12 % for 5 to 50 anonymity levels for IPUMS dataset.

When mining large data set, evolutionary algorithms like GA find optimal data sets. Results reveal the algorithm being capable of finding an optimal/near optimal solution for varying k-anonymity model levels. Performance metrics used include average accuracy, precision and recall. The classification accuracy declines by 4.43% when anonymity level is 50. The classification accuracy achieved at different anonymity level reduces with increase in level of anonymity for the IPUMS 99 Dataset by 0.11% to 3.59% for anonymity levels varying from 5 to 50.

In this proposed work, SA with a GA is used to optimize the feature selection. Results show that the algorithm is able to find an optimal solution or near optimal solution for varying Levels of k in anonymity model. The performance metrics used are average accuracy, precision and recall. The result shows that accuracy decreases by 0.11% to 2.68 % for IPUMS dataset and by 4.19% for k=50 when compared with non-anonymized data for Mushroom dataset.