CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 CONCLUSION

Data mining technology attracted interest to identify patterns and trends in large data collections. Extracted knowledge, expressed as association rules, decision trees or clusters, allowed locating patterns/regularities hidden in data but facilitated decision making. But it is evident that data collection and analysis includes personal information which can violate individual privacy. As data mining techniques reveal critical information on business, compromising free competition, disclosures of confidential/personal information must be prevented in addition to sensitive knowledge in a given context. Privacy protection in data mining is thus a crucial issue that has captured researchers’ attention.

Hence, research was devoted to addressing privacy preservation in data mining resulting in many data mining techniques which included privacy protection mechanisms based on various approaches. Various sanitization techniques were proposed to hide sensitive items/patterns based on removing reserved information/inserting noise in data. Privacy preserving classification methods prevent miners from constructing classifiers capable of predicting sensitive data. Also, recently proposed privacy preserving clustering techniques distort sensitive numerical attributes but preserve general features for cluster analysis.
Privacy-preserving data mining’s basic idea was extending data mining techniques to work with sensitive information masked modified data. What was at issue here was how to modify data and how to recover data mining result from it. Solutions were linked to data mining algorithms under study. This work investigated anonymization effect due to k-anonymity on the data mining classifiers. Data is anonymized for different granularity. 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.

K-anonymity method with differing k-levels was used. When mining large data set, evolutionary algorithms like GA find optimal data sets. Mushroom data sets evaluated the experiment and performance parameters like accuracy, precision, and recall, and were represented graphically with differing k-levels for granularity reduction. 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.

This work proposed a hybrid algorithm for privacy preserving in data mining. K-anonymity method with varying k-levels is used. Hybrid algorithms are used to overcome the shortcomings of the individual optimization techniques. In this work, evolutionary algorithms such as genetic algorithm and simulated annealing are used to find optimal data set. Mushroom data set has been used for evaluation, and the performance parameters like accuracy, precision and recall were represented graphically.
with varying k-levels for granularity reduction using k-anonymity. Experimental results demonstrate that with the increase in the k-anonymity levels, the performance of the classifier decreases but within an acceptable level.

The proposed Hybrid optimization is compared to other techniques in literature. Table 6.1 reveals classification accuracy achieved.

**Table 6.1 Comparison of the proposed hybrid optimization method with works available in the literature**

<table>
<thead>
<tr>
<th>K anonymity level</th>
<th>Proposed Hybrid optimization</th>
<th>Slava et al</th>
<th>Friedman et al</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>94.76</td>
<td>85.73</td>
<td>82.5</td>
</tr>
<tr>
<td>15</td>
<td>94.32</td>
<td>85.61</td>
<td>82.6</td>
</tr>
<tr>
<td>20</td>
<td>93.88</td>
<td>85.48</td>
<td>82.7</td>
</tr>
</tbody>
</table>

From Table 1 it can be seen that the proposed technique improves the classification accuracy. Classification accuracy is improved for Proposed Hybrid optimization by 9.52% and by 12.93% when compared to Slava et al and Friedman et al when k anonymity level is 10.

**6.2 FUTURE WORK**

- Performance of the proposed method with large number of instances to be investigated.
- Similar methodology to be applied for other datasets.