CHAPTER 5

CONCLUSION AND FUTURE SCOPE
5.1 Conclusion

The aim of the present investigation is to design an efficient model for privacy preserving classification of data streams. Studies were conducted to evaluate the performance of various anonymization techniques, classification techniques and anonymized data stream classification. It has been observed that:

- Amongst various anonymization techniques tested, (such as k-anonymity, l-diversity, population uniqueness and average re-identification risk), population uniqueness and k-anonymity preserved the privacy with least information loss.
- Amongst various classification techniques applied on data streams, maximum accuracy was achieved from J-48.
- Upon classifying the anonymized data streams, maximum privacy (41%) and accuracy (82.75%) has been obtained from J-48 when applied on the privacy preserved data stream using population uniqueness and k-anonymity. The accuracy achieved before and after anonymization was comparable.

Attempts have also been made to study the data stream challenges like concept drift and infinite length. It has been observed that:

- Utility of the data suffers when there is an attempt of improving the privacy of data stream. Utility and privacy depends on the dataset and the privacy preservation technique applied. Further, the accuracy and privacy changes with the increase in the size of the dataset.
- In synthetic data stream consisting of concept drift, the ensemble classifiers AUE and AWE provide better accuracy and consumes less memory in minimum response time.
- In real data stream OZABAGASHT and OZABAG classifiers provides maximum accuracy, consumes less memory in minimum response time and high kappa statistics. Ensemble classifiers provide better accuracy as compared to the single classifiers. Further, the performance of classifiers depends on type of data such as numerical, categorical, alpha numerical and mixed data type.
In an attempt to design an efficient model for privacy preserving classification of data streams, an efficient technique based on hashing, is designed and implemented on synthetic and real data streams. The privacy preserved data stream using this proposed technique is then classified applying various classifiers.

Hashing maps large data set of variable length to smaller data sets of fixed length and provides more reliable way of data retrieval. Time complexity of an algorithm is commonly expressed using big O notation. For hashing in best case time complexity is O(1) and in average case O(n). Hash value can be used to uniquely identify secret information. Data can easily be stored in secure binary format at local file system and vice versa which overcome memory issue. Hence, it has been observed that:

- The technique using hashing when compared with anonymization technique has shown improved privacy by 14.5%.
- When ensemble classifiers were applied on original data stream, OZABAGASHT provided maximum accuracy (97.1%), in minimum response time. When classification is applied on proposed hashing based privacy preserved data stream, maximum accuracy (96.40%), is obtained by OCBOOST classifier. It produces comparable accuracy.
- The increase in number of sensitive attributes has no effect on privacy achieve from hash based technique, whereas in the case of anonymization as the number of attributes increases the privacy reduces.

It is concluded that the model accurately classifies the privacy preserved data streams applying the new privacy preservation technique. The privacy achieved through the hashing based technique efficiently preserves the sensitive attributes with minimum information loss. The proposed model compensates for taking more execution time by providing the benefit of recovering the original data stream without information loss.
5.2 Future Scope

Besides the classification of privacy preserved data stream, it is also desirable to combine the classification with the privacy preservation technique. In future, efforts will be made to reduce the execution time of the designed technique without compromising the accuracy and privacy achieved. Further, enhancement is suggested to improve the technique to make it applicable for cloud based environment. This will enable multiple users to access simultaneously, as most of the data streams are generated through online applications.