Chapter 6

Conclusions and Future Work

6.1 Conclusions

This thesis presented the improvements of semi-supervised clustering technique that can improve the accuracy and reduce computational time. In this work the mainly two well-know methods are addressed viz. single-link and density-based clustering approaches that can solve real-world problems. The novelty of this thesis, introduced leaders concept that identify highly influenced patterns from the datasets. The advantage of this approach is to reduce the redundant patterns yield to reduce the running time of algorithms.

Constrained leader-based SSSL clustering performs with much reduced time when compared with SSSL clustering method with decrease in accuracy. The decrease in rand-index value for both synthetic and standard datasets with minimum of 0.1 to maximum of 0.3 clustering result due to the approximation error rate during the clustering process. Stopping condition is based on the no of classes in the dataset.

The clustering result is almost similar to the DBSCAN and C-DBSCAN but not exact. The $\epsilon$ and minpts are mandatory user input parameters. Achieving exact clustering result with much reduced time is one of challenges and also automatic finding of density based clustering method without $\epsilon$ and minpts is one of most challenging future research directions.
This thesis has developed a new approach leader based SSSL and Density based clustering for improving the classifier performance on IDS/IPS by introducing leaders concept that can work with less training examples with labeled and unlabeled records with Single-link clustering approach because it is very efficient and utilizes very limited resources. On the other hand DBSCAN is accurately detecting intrusions in less responsive time.

The experiments demonstrated in this thesis is reduced the classification accuracy by introducing leaders concept in both single-link and DBSCAN with clustering mechanism that lead to misclassification rate. It is noticed that our methodology is an effective way to improve the classification accuracy even with less number of training patterns with partial labeled samples.

The classifier is performed better when incorporating the unlabeled patterns with their predicted labels into the original training set. In this paper, we reported two-class problem, i.e., normal and anomaly. Further, this thesis work will continue towards applying more semi-supervised clustering approaches to improve the effectiveness of IDS/IPS for detecting multiple types of attacks.

6.2 Future work

In this thesis, with all the clustering methods used in the experimental studies, the number of clusters is given as the stopping criterion. With sufficient number of constraints, the proposed SSSL method can find a clustering result without the need for the user to give any stopping condition. Verifying this claim is one of the future directions of research.

Constrained leader-based SSSL clustering performs with much reduced time when compared with SSSL clustering method with decrease of accuracy. The de-
crease of rand-index value for both synthetic and standard datasets with minimum of 0.1 to maximum of 0.3 clustering result due to the approximation error rate during the clustering process. Stopping condition is based on the no of classes in the dataset. Leaders algorithm can be extended to finer level to eliminate approximation error and getting similar clustering result for Constrained l-SSSl and with SSSL method are the new future directions of research work.

The epsilon and Minpts are mandatory user input parameters. Achieving exact clustering result for Leader-based constrained DBSCAN and C-DBSCAN method, with much reduced time is one of challenge and also automatic finding of density based clustering method without epsilon and minpts is one of most challenging future research directions.

In this thesis, the proposed methods addressed the two-class problem, i.e., normal and anomaly in IDS/IPS. Further, this thesis research work will continue towards applying more semi-supervised clustering approaches to improve the effectiveness of IDS/IPS for detecting multiple types of attacks and also can be applied to any real world application such as Medical diagnosis in health care systems by using pairwise constraints, segmentation of customers in Marketing etc.