Chapter 8

Discussion and Conclusions

The principle theme of this research work is the study of pattern matching algorithms for Intrusion Detection Systems (IDS). For this, number of different models and learning approaches have been considered and investigated. This work has presented an array of existing machine learning algorithms for pattern matching and explored their characteristics. In this work it has been attempted to expedite the applicability of the solutions of pattern matching within the context of Network Intrusion Detection Systems. This work has also provided background and clear explanations of the algorithms.

In particular, we pursue research in unsupervised, supervised and semi-supervised approaches and elucidate these approaches. Furthermore, although the supervised algorithms are used directly in Intrusion Detection System (IDS) algorithms, its consideration as a sub-component to semi-supervised approach is a novel contribution within the IDS. Initially we investigate SOM algorithm in our research work as unsupervised approach. But unsupervised approach has its own drawback so we implemented supervised approach. Again, to our knowledge from literature survey, the performance of decision tree and SVM algorithms individually used for IDS is not up-to the mark. So, the combined approach of these algorithms has received much attention to improve the performance of IDS. Therefore, in our research work of supervised approach, we investigate the hybrid model of decision tree and Support Vector Machine (DT-SVM). The performance of this hybrid model, DT-SVM is poor for some attacks, so we decided to implement another supervised algorithm.

From literature survey for pattern recognition image processing, AdaBoost seems to be a good choice for supervised IDS. Because AdaBoost algorithm is: simple and easy to program, no parameters to tune (except T), provably effective, provided can consistently find rough rules of thumb. AdaBoost is less sensitive to noise that is added to the features space at some dimensions at the training phase. AdaBoost's classification complexity is
linear in the number of features used whereas SVM's classification complexity is much more expensive. This attracts us to use this algorithm for pattern matching IDS.

The motivation for applying the AdaBoost algorithm includes: 1) The AdaBoost algorithm is one of the most popular machine learning algorithms. Its theoretical basis is sound, and its implementation is simple. It has been applied to many pattern recognition problems, such as face recognition. However, the application of the AdaBoost algorithm to intrusion detection has not been explored so far. 2) The AdaBoost algorithm corrects the misclassifications made by weak classifiers, and it is less susceptible to over-fitting than most learning algorithms. Pattern recognition performances of the AdaBoost-based classifiers are generally encouraging. 3) Data sets for intrusion detection are a heterogeneous mixture of categorical and continuous types of features. The different feature types in such data sets make it difficult to find relations between these features. By combining the weak classifiers for continuous features and the weak classifiers for categorical features into a strong classifier, the relations between these two different types of features are handled naturally, without any forced conversions between continuous and categorical features. 4) If simple weak classifiers are used, the AdaBoost algorithm is very fast. So in our research work of supervised approach, we propose and investigate DS-AdaBoost algorithm. The contribution of this work is a new improved DS-AdaBoost supervised learning algorithm for Pattern Based Network Security.

To improve the performance of supervised approach we propose and investigate the semi-supervised approach. Another contribution of this work is a novel approach for Pattern Based Network Security using semi-supervised self learning is presented. This algorithm called SLA, which uses any supervised algorithm as a black box. This approach helps to reduce the dependability on the large scale labelled data. We implement and examined our semi-supervised approach using DS-AdaBoost supervised algorithm.

In presenting all the algorithms and their characteristics we also compare and contrast the algorithms in general, and also with a focus on the context of pattern matching in IDSs. By examining the requirements for IDSs we have revealed which of the available traits provided by pattern matching algorithms are most desired and which algorithms are indeed suited to fulfil the requirements.
Covering this broad range of pattern matching algorithms we discovered that the previous implemented approaches are lagging in accuracy. Specifically, our algorithm DS-AdaBoost, shows promise because it was able to outperform as a supervised as well as semi-supervised approach.

An attempt has been made to work upon the KDDCup99 and NSL-KDD data sets. The methods used are initially applied on KDDCup99 data set and then same procedure is used for the NSL-KDD data set. Initially, unsupervised and supervised approaches are considered for the classification. To improve the performance of IDS, semi-supervised approach is used for classification. At the time of performance investigation for such type of classification problem, the use of DS-AdaBoost and the hybrid approach of decision tree-SVM are found particularly attractive approaches. In which, DS-AdaBoost shows 99.96% and DT-SVM shows 99.88% training accuracy for supervised approach. The detection rate for DS-AdaBoost algorithm is 99.70% with false positive rate 0.06% for supervised approach and detection rate is 99.96% with false positive rate 0.055% for semi-supervised approach for KDDCup99 data set. For NSL data set the detection rate are 99.84 and 99.86 for supervised and semi-supervised approaches respectively. For NSL-KDD data set the false positive rate are 0.089 and 0.076 for supervised and semi-supervised approaches respectively. The performance of supervised DS-AdaBoost is good due to: use of decision stump with AdaBoost algorithm and the use of threshold to handle the problem of over-fitting. The result of semi-supervised DS-AdaBoost is improved due to the selection of most confident data based on entropy calculation ans statistical model. In view of these results, detailed optimization of various models such as AdaBoost-M1, J48, SMO, Naïve Bayes, and Bagging algorithms available in WEKA have been performed. The performance of Naïve Bayes algorithm is poor as compare to other algorithms. But using semi-supervised approach the performance of supervised algorithms can improve. From the result we can conclude that using NSL-KDD data set we can get better detection rate as compare to KDDCup99 data set. But the performance for semi-supervised approach using KDDCup99 data set is better than NSL-KDD data set. This is due to the variation of classes selected in the most confident data.

The key observations made from the results are highlighted below:
1. The experimentations show that the accuracy of SOM algorithm is very poor particularly for R2L, U2R, and Probe. Also SOM is the most time consuming algorithm.

2. The experimentations show that the performance of DT-SVM hybrid model is fine for Normal packets, DOS, U2R and Probe where the accuracy ranges from 94.45% to 100%, however for R2L, the accuracy ranges only from 68% to 84% maximum. These results show that the accuracy of even supervised DT-SVM hybrid model is not suffice to give the best performance. There is a need to improve the accuracy for R2L attacks. Even the accuracy for Normal packets, DOS, U2R and Probe are not good than some existing algorithms.

3. In pattern based network security, DS-AdaBoost has displayed better performance for both data sets using supervised and semi-supervised approaches.

4. Classification accuracy and performance of the algorithm can be application or data set dependent (i.e. data set size, features selected, data collection, the way the data is split between training and testing set etc.), but using DS-AdaBoost with supervised and semi-supervised approaches, perform equally well for KDDCup99 and NSL-KDD data sets.

5. A comparison between DS-AdaBoost, AdaBoost M1, Naïve Bayes, J48, Bagging, and SMO suggest that DS-AdaBoost has good performance for large and unbalanced data sets as compared to other classifiers, which showed that the use of more complex classifiers are not necessary for IDS.

6. While experimentations, it is noted that Naïve Bayes algorithms get failed due to the complexity of data set and it requires more time to extract the rules. DS-AdaBoost algorithm speed up the training performance time in many situations. Also, such type of classification problem contains numerical approach (like DS-AdaBoost) is the best suitable for large data set and improving generalization performance of the model.

7. The proposed Semi-supervised learning algorithm (SLA) gives noticeable enhancement in performance of the supervised approach. For supervised DS-AdaBoost detection rate is 99.7 which is improved to 99.96% using SLA. For supervised DS-AdaBoost false positive rate is 0.06% which is reduced to 0.055%.
This shows that the performance of supervised Ds-AdaBoost algorithm is improved using proposed semi-supervised SLA approach. Also the comparison of SLA with state-of-art algorithms from literature shows that performance of SLA is better than other algorithms.

DS-AdaBoost is found to be a good solution for IDS. The approach suggested based on Pattern Based Network Security renders better performance. This study shows that, any supervised algorithm can be used as a black box for semi-supervised learning approach. It helps to improve the performance of supervised learning algorithms. The contribution of this work is the Dynamic Threshold based, Semi-supervised learning algorithm that improved overall IDS accuracy, reduce FPR and improve the detection rate. Further this algorithm and methodology can be used in real life network environment.

8.1 Future Scope
This work lays the basis for many areas of future work. More classifiers can be implemented and tested on various data sets for IDS. There is a scope for detecting more attacks and can be implemented on larger real data size. There is a scope for improvement of the performance and accuracy of semi-supervised approach using various techniques for selecting most confident data. Here is a lot of potential to improve the performance of semi-supervised DS-AdaBoost for structured prediction problems in robotics, computer vision and natural language processing. This requires the integration of various methods for selecting the most confident data to our current framework. One may use weight of each instance or information gain or the confidence to select the most confident data for semi-supervised approach. More classifiers can be implemented and tested on various data sets for wireless network IDS.