CHAPTER 6

CONCLUSION

6.1 SUMMARY

It was proposed to study the effect of accuracy of classification on legacy modules, using various data mining techniques inference for practical reliability assessment.

Experimental analysis was carried out on the KC3 dataset using Naïve Bayes classification and K nearest neighbor classification. LOC and halstead attributes were used for analysis in the KC3 dataset. Using XLMiner the data was partitioned into training data and validation data in the ratio 40:60. NB classification and K nearest classification algorithms were applied. The percentage error on classification in Naïve Bayesian was 17.87% and 22.81 % for K nearest neighbor with K=1.

Based on the experimental analysis of KC3 dataset for which the HEIF embedded in SRBIT model was applied. The data preparation was in the same format as in the previous research. Using WEKA for classification it was found that performance of the K nearest neighbor classification algorithm improved by 5.42% for K=1.

In the previous experimental analysis, it was seen that the error classification percentage improved if the data is preprocessed using immune technique methods. Instead of using the regular halstead and LOC attributes, it was proposed to use the density attributes in the KC3 dataset. The error density and the cyclomatic density attributes were selected. Each record was
analyzed and data distribution was normalized. The data preparation was standardized with 40 percentage of the records being training data and the rest being validation data. Neural network classifier algorithm was applied on the records and the classification error was 0.38 percentage. There is a very good improvement in the error classification compared to the previous methods.

From the above experimental analysis, it can be inferred that data mining techniques with data pre-processing is a very promising area which can help the IT managers in maintaining legacy software. It also provides a excellent method for the IT managers to plan resources for improving erroneous modules using classification technologies which can classify modules with a high accuracy. The amount of digital data / size of databases has been exploding during the past decade, while the number of analysts available to analyze the data has been static.

6.2 SCOPE FOR FURTHER RESEARCH

Our research has shown that algorithms in pre-processing of data can improve the error classification using neural network, NB and K nearest neighbor algorithm. Further, research could be established in the areas of ensemble learning, Graphical Models and Hierarchical Probabilistic Representations, Large scale optimization to further improve the classification error rate. Research in the areas of new fundamental problems of data mining is crucial and these challenges can be grouped into the following broad challenges:

- Creating algorithms and systems for massive and high dimensional dataset
- Creating algorithms and systems for new types of data
- Creating algorithms, protocols for distributed data
- Creating appropriate privacy and security models for data mining.