Chapter – 7

Conclusion and Future Work

This chapter summarizes the research work carried out in the thesis. The objective of this work is to propose an integrated model using clustering and classification techniques for providing a solution to the basic machine learning problems viz. improper (noisy, missing and incomplete data), Imbalanced, and high dimensional data.

This chapter comprises of two sections, the first section concludes the findings derived through experiments conducted on several benchmark datasets taken from UCI machine learning repository and the second section suggests the scope of future enhancements.

7.1 Conclusion

The research carried out here provides a generic and logical solution to the above-mentioned problems and dealt with separately in chapters. The models are presented, evaluated, and compared for each of the issues on various performance measuring parameters with State-of-the-art methods.

The first proposed model presented in chapter-4 provides a Cluster-based approach using an under-sampling solution to balance the Imbalanced data. A study for binary class distribution on 12 data sets openly available in UCI machine learning repository with different degrees of imbalance nature has been conducted and presented. The model is capable of handling between-class Imbalance distribution and within-class distribution .it can also handle different degrees of imbalanced distribution and finally, it balances the imbalanced data and that can be applied to any traditional classifier. The proposed model is simple yet effective in order to classify the Imbalanced distribution of Data.

The second proposed model which is presented in chapter-5 is applicable as a Feature compression and Extraction technique to build a better feature space. The experiments were conducted on nine-benchmark dataset taken from the UCI machine repository with diverse degrees of dimensionality. The Comparative Analysis of the Proposed Approach with a standard correlation-based Feature Selection approach (RELIEF and Info-Gain Approach) was presented on 20, 40, 60, 80, and 100 % of the features on three performance measuring parameters; Accuracy, F-Measure, and ROC. The centroid of each cluster has been taken as the individual cluster representative. It reduces dimensionality, complexity, and computation time, and increases comprehensibility and the overall performance of classification algorithms. Therefore, the proposed model first identifies the relevant features that may lead to accurate results.

The third proposed model presented in chapter-6 is to improve the performance of any classifier using a hybrid model (prior clustering to classification). The research experiments observed that the hybrid model is more refined and accurate than a single individual classifier. The objective is to utilize the strength of one method to complement the weaknesses of another. If we are interested in the best possible classification accuracy, it might be difficult to find a single classifier that performs as good as a hybrid model. The experiment conducted on 13-benchmark dataset taken from the UCI machine learning repository. The results evidently revealed that this integration process generates a more precise and accurate model.

7.2 Future Enhancements

When it comes to future enhancements discussion of this thesis work, a lot of continuation could be derived.

In the first model, the experiments are carried out on binary classification problems only the work can be extended for multiclass classification problems also.

In the second model only centroid based cluster representative selection methods are considered the work can be extended to examine other available methods viz. random selection and ranking based method.

In the third model the performances are tested on only the Support vector machine (SVM) classifier, the work can be extended by testing the model’s performance on other classifiers also.