Abstract

Data is increasing at an unimaginable rate every year. The area of Data Mining has arisen over the last decade to address this problem. Progress in digital data acquisition and storage technology has resulted in the growth of huge databases. This has occurred in all areas starting from simple applications like supermarket transactions, railway reservations to the more complex and complicated ones like space research, molecular databases, images, astronomical bodies, etc. Using this data to discover hidden knowledge, unexpected patterns and unknown information is Data Mining.

Data Mining includes two indispensable tasks, Clustering, and Classification. The integration of these tasks together can undertake challenging machine-learning problems. Taking advantage of these methods is a significant research area. There are few quality issues present in most of the real-world datasets that negatively influence the performances of the Classifier they are imbalanced, high dimensional, noisy and incomplete Data.

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

The first proposed model 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 that presented a three-fold solution. Firstly, it balances imbalanced data using the K-means clustering approach. The main purpose of this approach is to selectively discard

majority instances from the dataset to make the distribution balanced, which can be applied to any traditional classifier. Secondly, it is capable in handling between class imbalance distribution and within-class distribution. Thirdly, it can handle the different degrees of imbalance distribution. The proposed model is simple yet effective in order to classify the imbalanced distribution of data.

The second proposed model is applicable as a Feature Compression and Extraction technique to build a better feature space for machine learning applications. High dimensional data generally diminish the accuracy and efficiency of Data Mining algorithms. The advantage of the proposed model is; it first identifies the relevant features that may lead to accurate results and then classifies it. The experiments were conducted on 9 benchmark datasets 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 these datasets.

The third proposed model is to improve the performance of any classifier using a hybrid model (prior clustering to classification). The empirical results prove 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 the interest is to find the best possible classification accuracy, it might be difficult to find a single classifier that performs as good as the proposed hybrid model. The experiment has been conducted on 13-benchmark datasets taken from the UCI machine learning repository. The results evidently revealed that this integration process generates a more precise and accurate model.

The developed models have been published in Conference Proceedings /International Journals and the details of the publications are mentioned in the end.