ABSTRACT

In supervised classification, a new and burning challenge has emerged for researcher community in data mining is Class Imbalance Learning (CIL). The problem of class imbalance learning is of great significance when dealing with real-world datasets. The data imbalance problem is more serious in the case of binary class, where the number of instances in one class predominantly outnumbers the number of instances in another class. Class imbalance learning is one of the issues which degrade the performance of any traditional classifier. To address these issues of class imbalance problem the present work proposed various novel and effective class imbalance learning (CIL) algorithms.

The present study proposed a practical algorithm to deal with the data imbalance classification problem when there are datasets with different imbalance ratios. A new algorithm called Prominent Recursive Minority Oversampling Technique; PRMOTE is proposed. We proposed to recursively oversample the most prominent examples in the minority set for handling the problem of class imbalance. So far no researcher used only novel instances subset in minority subset for oversampling in developing any schemes for solving the problem of class imbalance. The proposed framework is implemented in two stages.

At first stage, the dataset is applied to one of the base algorithm for preprocessing (removing mostly misclassified instances). On the next phase the only novel instances in the majority subset are selected for synthetic oversampling. In the final stage minority subset and the improved majority subset with prominent instances are combined to form strong dataset and is applied to a base algorithm for performance evaluation. The novelty of the proposed PRMOTE is also investigated by implementing a variation of PRMOTE using C4.5, Naïve Bayes, Logistic Regression and Support Vector Machines as the base algorithms.
In total one frame work and set of four novel algorithms are proposed as the contributions of this thesis along with this a case study is conducted for more elaborate technical analysis. The set of novel contributions are dubbed as PROMOTE-C4.5, PROMOTE-NB, PROMOTE-LR and PROMOTE-SVM with C4.5, Naïve Bayes, Logistic Regression and Support Vector Machines as the base algorithms respectively.