Clustering Defaults into Different Groups to Improve Default Model Fit and Predictive Performance

Key Words: Classification, Class Imbalance, Logistic regression, EM algorithm

In consumer credit, it is common to encounter two-class classification tasks with highly imbalanced data sets in which one class is very rare compared to the other; (e.g. default or fraud group). Owen (2007) demonstrated, asymptotically, that logistic regression only uses the rare class data via the rare class mean vector, which means the broader distributional structure of the minority class would not be taken into account by logistic regression. Relabeling the minority class data into distinct groups is a potential approach to alleviate this problem, which could be implemented in various ways. Practically, relabeling faces a computational burden, because it is an optimization over a discrete set and hence faces a combinatorial explosion. To manage this computational burden, we propose a novel procedure using the Expectation-Maximization algorithm and multinomial logistic regression to evaluate the multi-class relabeling. In our simulation study, we compare this EM procedure with a brute force method (Genetic Algorithm) and find EM provides a satisfactory and faster performance. We illustrate applications of this method using several real credit data to demonstrate its effectiveness. Our results also find that relabeling could provide meaningful relabeled minority clusters showing the difference in characteristics between different default groups.

Reference: Owen, A.B., 2007. Infinitely imbalanced logistic regression. Journal of Machine Learning Research, 8(Apr), pp.761-773.