Solving class imbalance problem using bagging, boosting techniques, with and without using noise filtering method

Article type: Research Article

Authors: Rekha, G. (https://content.iospress.com:443/search?q=author%3A%28%22Rekha, G.%22%29)[a] | Tyagi, Amit Kumar (https://content.iospress.com:443/search?q=author%3A%28%22Tyagi, Amit Kumar%22%29)[b;*] | Krishna Reddy, V. (https://content.iospress.com:443/search?q=author%3A%28%22Krishna Reddy, V.%22%29)[a]

Affiliations: **[**a**]** Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Guntur 522502, India | **[**b**]** Lingaya's Vidyapeeth, Faridabad 121002, India

Correspondence: [*] Corresponding author: Amit Kumar Tyagi, Lingaya's Vidyapeeth, Faridabad 121002, Haryana, India. Tel.: +91 9487868518;E-mail: amitkrtyagi025@gmail.com (mailto:amitkrtyagi025@gmail.com).

Abstract: In numerous real-world applications/domains, the class imbalance problem is prevalent/hot topic to focus. In various existing work, for solving class imbalance problem, almost data is labeled as one class called majority class, while fewer data is labeled as the other class, called minority class (more important class to focus). But, none of the work has performed efficiently (in terms of accuracy). This work presents a comparison of the performance of several boosting and bagging techniques from imbalanced datasets. The wide range of application of data mining and machine learning encounters class imbalance problem. An imbalanced datasets consists of samples with skewed distribution and traditional methods show biased towards the negative (majority) samples. Note that popular pre-processing technique for handling class imbalance problems is called over-sampling. It balances the datasets to achieve a high classification rate and also avoids the bias towards majority class samples. Over-sampling technique takes full minority samples in the training data into consideration while performing classification. But, the presence of some noise (in the minority samples and majority samples) may degrade the classification performance. Hence, the work presents a performance comparison using boosting and bagging (i.e., with both techniques) with and without using noise filtering. This work evaluates the performance with the state of-the-art methods based on ensemble learning like AdaBoost, RUSBoost, SMOTEBoost, Bagging, OverBagging, SMOTEBagging on 25 imbalance binary class datasets with various Imbalance Ratios (IR). The experimental results show that our approach works as promising and effective for dealing with imbalanced datasets using metrics like F-Measure and AUC.

Keywords: Class imbalance problem, ensemble learning method, noise filter, boosting, bagging

DOI: 10.3233/HIS-190261

Journal: International Journal of Hybrid Intelligent Systems (https://content.iospress.com:443/journals/international-journal-of-hybrid-intelligent-systems), vol. 15, no. 2, pp. 67-76, 2019

Published: 13 June 2019

Price: EUR 27.50