Impact of Imbalanced Datasets on Classification Algorithms: A Comparative Study

Expose

This research paper investigates the impact of imbalanced datasets on classification algorithms, focusing on Support Vector Machine (SVM), Decision Tree, Logistic Regression, and Naive Bayes. Imbalance learning is then explained, outlining various methods such as oversampling, under sampling, and hybrid approaches. The working principles of SVM classification are elucidated, followed by an analysis of how imbalanced datasets affect SVM performance. The challenges arising from imbalanced datasets are further explored by repeating the analysis for Decision Tree, Logistic Regression, and Naive Bayes algorithms.

To address the imbalance issue, different sampling techniques are considered. The suitability of each sampling method for specific machine learning algorithms is discussed, considering their characteristics and limitations. Additionally, the study examines the acceptable range of class quantity differences that can affect different algorithms, providing insights into the reasons behind these disparities.

This research contributes to a better understanding of the impact of imbalanced datasets on classification algorithms and provides insights into suitable sampling techniques for different algorithms. The findings aim to guide practitioners in selecting appropriate algorithms and methodologies to improve the accuracy of classification tasks in the presence of class imbalance.

Dataset:

1. Is this a good Customer: This dataset has two class labels, yes or no. It has 13 features and one target attribute. (https://www.kaggle.com/datasets/podsyp/is-this-a-good-customer)

2. Wine Quality: This dataset has 11 class labels, 0-10. It has 12 features and one target attribute.

(https://www.kaggle.com/datasets/rajyellow46/wine-quality)

References:

1. Rehan Akbani, Stephen Kwek, and Nathalie Japkowicz, "Applying Support Vector Machines to Imbalanced Datasets" 2004.

2. H. Zhang, Z. Li, H. Shahriar, L. Tao, P. Bhattacharya and Y. Qian, "Improving Prediction Accuracy for Logistic Regression on Imbalanced Datasets," 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), Milwaukee, WI, USA, 2019, pp. 918-919, doi: 10.1109/COMPSAC.2019.00140.

3. H. He and E. A. Garcia, "Learning from Imbalanced Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, Sept. 2009, doi: 10.1109/TKDE.2008.239.