MACHINE LEARNING ALGORITHMS FOR DATA LABELING: AN EMPIRICAL EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The lack of labeled data is a major problem in both research and industrial settings since obtaining labels is often an expensive and time-consuming activity. In the past years, several machine learning algorithms were developed to assist and perform automated labeling in partially labeled datasets. While many of these algorithms are available in open-source packages, there is no research that investigates how these algorithms compare to each other in different types of datasets and with different percentages of available labels. To address this problem, this paper empirically evaluates and compares seven algorithms for automated labeling in terms of accuracy. We investigate how these algorithms perform in six different and well-known datasets with three different types of data, images, texts, and numerical values. We evaluate these algorithms under two different experimental conditions, with 10% and 50% labels of available labels in the dataset. Each algorithm, in each dataset for each experimental condition, is evaluated independently ten times with different random seeds. The results are analyzed and the algorithms are compared utilizing a Bayesian Bradley-Terry model. The results indicate that while the algorithms label spreading with K-nearest neighbors perform better in the aggregated results, the active learning algorithms query by instance QBC and query instance uncertainty sample perform better when there is only 10% of labels available. These results can help machine learning practitioners in choosing optimal machine learning algorithms to label their data.

REFERENCES