UNIOESTE at LeQua 2024: Combining the top-ranked quantifiers

Luiz Luth, Daniel Ojeda, Guilherme Gomes, and Andre Maletzke

Western Paraná State University, Foz do Iguaçu, Brazil
{luiz.junior90, daniel.ojeda, guilherme.gomes, andre.maletzke}@unioeste.br

Abstract. Keywords: Counting · Class prevalence shift · label shift

1 Introduction

In recent years, the problem of class prior estimation (also known as quantification or learning to quantify) has gained significant attention due to its importance in real-world applications where estimating the distribution of classes in unlabelled data is crucial. The challenge extends beyond classifying individual instances to accurately predicting the proportion of different classes in a dataset, a task complicated by prior probability shift (or label shift), where the class distribution in the test data differs from the training data, making standard supervised learning approaches less effective.

The LeQua 2024 competition addresses these challenges through several tasks, and our proposal specifically aims to tackle Task T1. This task focuses on evaluating binary quantifiers, which are tasked with predicting the relative frequencies of a class and its complement under conditions of prior probability shift.

Our proposal for Task T1 of the LeQua 2024 competition addresses the challenge of predicting class proportions under prior probability shift using a straightforward and effective strategy. Instead of relying on a single quantifier, we propose to run several quantifiers on the target dataset. Each quantifier generates a prediction for the class proportions, and we then rank these quantifiers based on their performance. Once ranked, we select the top-performing quantifiers and form the final prediction by averaging the predictions of the top-ranked models. This approach leverages the strengths of multiple quantification methods, increasing robustness and reducing the reliance on any one model. By combining the predictions of the best-performing quantifiers, we aim to produce more reliable and accurate estimates of the class proportions, effectively tackling the prior probability shift that characterizes Task T1.

2 Method

This section describes the method applied to obtain predictions for Task T1. First, we evaluated a set of base classifiers for Task T1, including CatBoost [1], Random Forests [2], Support Vector Machines [3], and XGBClassifier [4]. We conducted a grid search for each classifier to determine the best hyperparameters, as detailed in Table 1.

Table 1. Parameter grids for different models

Model	**Parameter**	**Values**
XGBoost	`n_estimators`	{100, 200, 300}
	`gamma`	{0.01, 0.1}
	`max_depth`	{3, 6, 9, 12}
	`learning_rate`	{0.001, 0.01, 0.1, 1}
CatBoost	`depth`	{4, 5, 6, 7, 8, 9, 10}
	`learning_rate`	{0.01, 0.05, 0.2, 0.5, 0.8, 1.0}
	`iterations`	{10, 20, 50, 100}
RandomForest	`n_estimators`	{100, 200, 300, 400, 500, 600, 700, 800}
	`criterion`	{`gini`, `entropy`}
	`max_depth`	{-1, 10, 50, 100}
SVC	`kernel`	{`linear`, `poly`, `rbf`, `sigmoid`}
	`gamma`	{`scale`, `auto`}

Furthermore, to account for label shift, we applied the correction technique proposed by Lipton et al. (2018) [5], which adjusts the predictions to mitigate the impact of prior probability shift in the test data.

Next, we employed several well-established quantifiers to generate the class proportion estimates. We selected the best-performing quantifiers for binary problems as described by Schumacher et al. (2021) [6]. We used the following quantifiers: DyS, HDy, SORD, CC, ACC, MS, and EMQ, each with their respective default parameters.

Finally, to generate the predictions, we ranked all the quantifiers and measured the Mean Absolute Error (MAE) for each individual. We also evaluated the performance by combining the quantifiers in groups of three and five, calculating the average prediction for each group. We then compared the performance of the individual quantifiers and their combinations, selecting the best-performing setup based on these results to predict the test set. This approach ensured that we identified the most robust configuration for accurate quantification under the conditions of prior probability shift.

3 Results

The results of our experiments indicated that the best-performing setup for Task T1 was a combination of both an individual classifier and a group of quantifiers. Specifically, the SVC classifier with `gamma` = `auto` and `kernel` = `poly` yielded the most accurate results among the base classifiers. In terms of quantifiers, the combination of DyS, HDy, and EMQ proved to be the most effective, providing the lowest Mean Absolute Error when their predictions were averaged. This combined approach outperformed other configurations, and thus it was selected to predict the test set, ensuring robust and accurate quantification under the conditions of a prior probability shift.

4 Conclusion

Our approach to Task T1 of the LeQua 2024 competition effectively addressed the challenge of prior probability shift by combining multiple quantification techniques and leveraging an optimized classifier. To further enhance our results, we applied a correction technique that adjusts predictions to mitigate the impact of prior probability shift in the test data, ensuring more reliable class proportion estimates. By averaging the predictions of the top quantifiers and applying the correction, we successfully selected the best-performing setup, which proved to be effective in handling distributional shifts. This strategy highlights the importance of combining classifier performance, quantification methods, and correction techniques to achieve high predictive accuracy under challenging conditions.

References

1. L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features, 2017," *arXiv preprint arXiv:1706.09516*, vol. 201, 2017.
2. L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.
3. C. Cortes, "Support-vector networks," *Machine Learning*, 1995.
4. T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
5. Z. Lipton, Y.-X. Wang, and A. Smola, "Detecting and correcting for label shift with black box predictors," in *International conference on machine learning*. PMLR, 2018, pp. 3122–3130.
6. T. Schumacher, M. Strohmaier, and F. Lemmerich, "A comparative evaluation of quantification methods," *arXiv preprint arXiv:2103.03223*, 2021.