An Evaluation of Methods for Unlabeled Class Prevalence Hypothesis Testing

David Kaftan[1,2], Anna Bershteyn[2], and Richard Povinelli[1]

[1] Marquette University, Milwaukee WI, USA
{david.kaftan,richard.povinelli}@marquette.edu
[2] NYU Grossman School of Medicine, New York NY, USA

Abstract. The machine learning task of quantification aims to estimate the proportion of classes in an unlabeled dataset. This task is made difficult by label shift, where the class distribution of the target data differs from the training data. While many methods focus on precise prevalence estimation, a common practical goal is simply to detect whether a class's prevalence has shifted from a nominal value. This study systematically compares four methods for detecting such deviation of class prevalence in binary classification tasks: a one-sample t-test, a one-sample KS-test, an Expectation-Maximization Likelihood Ratio Test (EM LRT), and a bootstrapped confidence interval approach (EM Bootstrap CI). Using an Artificial Prevalence Protocol – a procedure for varying prevalence in sampled test sets – on two real-world and one simulated dataset, we evaluated the methods based on their calibration and discriminatory power. Our findings indicate that calibration of the underlying classifier is critical for the EM-based methods to achieve a target 5% false positive rate (FPR). The EM LRT, when paired with a calibrated classifier, consistently maintained an FPR near the target and had the highest discriminatory power as measured by the area under the receiver operating characteristic curve. In contrast, the t-test proved to be robust to calibration status of the underlying classifier, though it had discriminatory power lower than the EM LRT and EM Bootstrap CI. This empirical evaluation provides practical guidance on the performance characteristics of different unlabeled class prevalence hypothesis tests, highlighting the EM LRT discriminatory power and the importance of calibration for likelihood-based approaches.

Keywords: Hypothesis Testing · Quantification · Detection.

1 Introduction

The machine learning field of quantification, also known as class prevalence estimation or learning to quantify, focuses on accurately estimating the proportion of each class within a given unlabeled dataset. While precise point estimates of class prevalence are often the direct goal of quantification, there are numerous practical scenarios where a slightly different objective comes to the forefront: detecting whether a class's prevalence differs significantly from a predefined,

nominal value. For instance, in public health, a health authority might not need to know the exact percentage of a population affected by a certain disease, but rather whether its prevalence has crossed an epidemic threshold [13]. Similarly, in quality control, a manufacturer might simply need to know if the defect rate in a batch of products has changed from a historical norm. We refer to this task of detecting changes in class prevalence as unlabeled class prevalence hypothesis testing.

In this paper, we investigate unlabeled class prevalence hypothesis testing as a one-sample, two-tailed hypothesis testing problem. We systematically compare methods for detecting when the prevalence of a binary class significantly deviates from a nominal value in simulated and benchmark datasets.

2 Background

The background section is split into two subsections. First, we will discuss prior work in dataset shift detection. Second, we introduce the task of quantification, including likelihood ratio testing methods for label shift detection and methods for constructing confidence intervals.

2.1 Dataset Shift Detection in Machine Learning

Dataset shift detection is a very similar task to unlabeled prevalence hypothesis testing. Dataset shift detection seeks to identify if a test set is sampled from a different distribution from the train dataset. Dataset shift between train and test sets can cause machine learning model performance to degrade and has therefore become a focus of machine learning researchers. Machine learning models can be designed to be robust to dataset shift, or dataset shift can be detected and models retrained. Rabanser et al evaluate several methods for detecting various types of dataset shift between train and test sets [20]. Rabanser took a two stage approach. First, the dimensionality of the train and test sets were reduced, then statistical tests comparing the two sets were applied. Several types of dataset shift were evaluated, including label shift, where prevalence of classes shifts between training and test sets. They found that Black Box Shift Detection [17] paired with a two-sample Kolmogorov-Smirnoff test (KS-test) performs consistently well at detecting various types of dataset shift, including label shift.

Unlabeled class prevalence hypothesis testing and label shift detection are related, but different tasks. Label shift detection seeks to detect if training and testing sets were sampled from populations with different class prevalence, while unlabeled class prevalence hypothesis testing detects if the prevalence in an unlabeled set deviates from a specific, nominal value.

2.2 Quantification

The goal of quantification is to estimate the prevalence of different classes in an unlabeled dataset, rather than predicting the class of individual samples.

Work on quantification has a long history in several fields, including epidemiology where researchers have long sought to estimate the prevalence of diseases given imperfect assays [14]. One of the key motivations behind quantification is to improve the accuracy of a classification model [9]. In these cases, classifier predictions are treated as posterior probabilities with respect to covariates and, therefore, a function of class priors. When a prior probability change - known as label shift - is detected, these posterior probabilities are adjusted to reflect the newly estimated prior. A common way to estimate a prior is to determine the prior probability that maximizes the likelihood of observing the unlabeled dataset (*eq 1*), where X is a sequence of unlabeled data (often the output of a classifier), and ω_i represents class i.

$$\begin{aligned} Likelihood\,(X) &= \prod P\,(x_i) \\ &= \prod_{i=1}^{N} P\,(x_i|\omega_0)\,P(\omega_0) + P\,(x_i|\omega_1)\,P(\omega_1) \end{aligned} \tag{1}$$

An expectation maximization (EM) algorithm to maximize likelihood with respect to prior probabilities was formalized by Saerens [21] (*eq 2*).

$$\begin{aligned} \hat{p}^{(0)}(\omega_i) &= \hat{p}_t(\omega_i) \\ \hat{p}^{(s)}(\omega_i|x_k) &= \frac{\frac{\hat{p}^{(s)}(\omega_i)}{\hat{p}_t(\omega_i)}\hat{p}_t(\omega_i|x_k)}{\frac{\hat{p}^{(s)}(\omega_0)}{\hat{p}_t(\omega_0)}\hat{p}_t(\omega_0|x_k) + \frac{\hat{p}^{(s)}(\omega_1)}{\hat{p}_t(\omega_1)}\hat{p}_t(\omega_1|x_k)} \\ \hat{p}^{(s+1)}(\omega_i) &= \frac{1}{N}\sum_{k=1}^{N}\hat{p}^{(s)}(\omega_i|x_k) \end{aligned} \tag{2}$$

where $\hat{p}^{(s)}(\omega_i)$ is the prior estimate at the s^{th} iteration of the expectation maximization algorithm, $\hat{p}^{(s)}(\omega_i|x_k)$ are the classifier outputs adjusted for the new prior estimate, and p_t denotes that under the training set. Expectation maximization is an algorithm which maximizes the likelihood of observing our dataset [8]. Though several methods for point estimates of binary class prevalence have since emerged [6], the EM approach has been shown to be difficult to beat, particularly when the underlying classifier is well calibrated [3, 10, 11].

Likelihood Ratio Tests for Detecting Label Shift Saerens et al. suggest using a likelihood ratio test to detect label shift [21]. They note that two times the log likelihood ratio of the unlabeled data under the true prevalence vs the maximum likelihood prevalence is distributed according to a chi-squared distribution. Label shift is then detected if the cumulative density function of the chi-squared distribution evaluated at two times the log likelihood ratio exceeds one minus the specified false positive rate (i.e., 0.05).

Confidence Interval Methods A fairly straightforward approach to unlabeled class prevalence hypothesis testing is to estimate confidence intervals around class prevalence estimates. The null hypothesis is then rejected when the prior associated with the null hypothesis falls outside of the confidence intervals. Confidence intervals are often constructed in quantification using standard bootstrapping techniques [16, 22]. Daughton and Paul developed the Error-Adjusted Bootstrap method to account for the errors associated with imperfect classifiers [7]. Tasche evaluated the coverage and confidence interval width for confidence intervals generated by bootstrapping several quantification methods [22]. Tasche found that the underlying classifier performance greatly impacted the confidence interval coverage and width.

3 Experiments

In this section, we describe the experimental protocols, datasets, and detection methods used.

3.1 Experimental Protocol

We run our experiments in an Artificial Prevalence Protocol (APP) [9] in the presence of label shift. An APP creates a variety of prevalences in the test set through random sub-sampling. We assumed that the class-conditional densities of classifier outputs are stationary between the train and the test set. We first split our training and testing datasets with a 50%-50% train-test stratified split. We train our classifiers on the entire training set. We construct our testing set by first selecting the number of positive instances to sample from a binomial distribution with positive class probability set to our artificially selected prevalence. To determine how performance scales, we use a test sample size of of 50, 75, and 100. We then sample, without replacement, the corresponding number of positive and negative samples from the test dataset.

We construct artificial prevalences under which the null hypotheses are true, and under which the true prevalence is 10% higher than that of the null hypothesis. We consider 3 null hypotheses: prevalence equal to 25%, 50%, and 75%. We calculate false positive rates (FPR) as the percent of simulations generated according to the null hypothesis that resulted in rejecting the null hypothesis. Similarly, we calculate true positive rates (TPR) as the percent of simulations generated with prevalence 10% higher than the null hypothesis that resulted in rejecting the null hypothesis. We simulate each scenario 2000 times to get stable estimates of performance. A summary of our experimental protocol can be found in Table 1.

3.2 Metrics

We measure the calibration and discriminatory power of each hypothesis test. A well-calibrated test has a false positive rate consistent with the significance level

of the test. In our experiment, we report the false positive rate when each test is evaluated at a significance level of 0.05, so tests with false positive rates close to 5% should be considered calibrated. We report discriminatory power of each test as the area under the curve of the receiver operating characteristic (AUC).

3.3 Datasets

We use three datasets in our experiments, including the Wisconsin Breast Cancer dataset [1], the Pima Indians Diabetes dataset [2], and a simulated dataset. All three datasets have binary classification targets. Datasets were chosen to have a breadth of input feature space. The Wisconsin Breast Cancer dataset has 30 features, the Pima Indians Diabetes dataset has eight features, and the simulated dataset has a single feature. The simulated dataset was generated by sampling a single feature from normal distributions ($\sigma = 1$) with zero mean for negative classes, and 0.85 mean for positive classes.

3.4 Detection Methods

Logistic regression classifiers are trained. We evaluate performance with and without calibration. We calibrate our regression model using Platt's method [19], which fits a logistic regression model of our classifier outputs to the labels across cross-validation folds. Our classifier outputs are then fed into four detection methods: t-test, KS-test, expectation maximization likelihood ratio tests, and bootstrapped confidence intervals around expectation maximization estimates.

T-Test T-tests are commonly used to detect differences in population parameters. We perform a two-sided, one-sample t-test, comparing the mean classifier probability output on the unlabeled test set to the expected value under the null hypothesis (population prevalence equals p_0). The expected value is computed according to *eq 3*.

$$E[X] = E[X|\omega_1]p_0 + E[X|\omega_0](1 - p_0) \tag{3}$$

KS-test Rabanser demonstrated that a two-sample KS-test applied to the output of a classifier can detect label shift from train to test distribution [20]. One-sample KS-tests can be used to test if a univariate sample comes from a specified known distribution by comparing the empirical cumulative density function (CDF) of the sample to the CDF of the known distribution. In this application, the sample of classifier outputs are compared to a mixture CDF of the class conditional distributions under the null hypothesis, defined by *eq 4*.

$$P_{H_0}(X \leq x) = P(X \leq x|\omega_1)p_0 + P(X \leq x|\omega_0)(1 - p_0) \tag{4}$$

Expectation Maximization Likelihood Ratio Test We implement the expectation maximization (EM) algorithm proposed by Saerens et al., [21] initializing the iterative algorithm according to our null hypothesis such that $\hat{p}^{(0)} = p_0$. Saerens proposes a log likelihood ratio comparing the likelihood of the maximum likelihood prior $\hat{p}^{(s)}(\omega_i)$ to the prior in the training dataset $\hat{p}_t(\omega_i)$. We modify the likelihood ratio to compare any prior p_0 to the maximum likelihood estimate by adjusting $\hat{p}_0(\omega_i|x_k)$ from classifier outputs $\hat{p}_t(\omega_i|x_k)$ according to *eq 5*.

$$\hat{p}_0(\omega_i|x_k) \;=\; \frac{\frac{\hat{p}_0(\omega_i)}{\hat{p}_t(\omega_i)}\hat{p}_t(\omega_i|x_k)}{\frac{\hat{p}_0(\omega_0)}{\hat{p}_t(\omega_0)}\hat{p}_t(\omega_0|x_k) + \frac{\hat{p}_0(\omega_1)}{\hat{p}_t(\omega_1)}\hat{p}_t(\omega_1|x_k)} \tag{5}$$

The χ^2 statistic is then calculated according to *eq 6*.

$$2log\frac{\prod_{k=1}^{N}\frac{\hat{p}^{(s)}(\omega_i)}{\hat{p}^{(s)}(\omega_i|x_k)}}{\prod_{k=1}^{N}\frac{\hat{p}_0(\omega_i)}{\hat{p}_0(\omega_i|x_k)}} \tag{6}$$

We then evaluate the χ^2 test with 1 degree of freedom. We refer to this test as the Expectation Maximization Likelihood Ratio Test (EM LRT).

Bootstrapped Confidence Intervals Finally, we create bootstrapped confidence intervals. We resample classifier outputs of the unlabeled test set with replacement, then run Saerens' expectation maximization algorithm to calculate our bootstrap prevalence estimate. The 2.5^{th} and 97.5^{th} quantiles of the bootstrapped prevalence estimates are taken as the confidence interval. If p_0 falls outside of the confidence interval, we reject the null hypothesis. We refer to this method as the EM Bootstrap CI.

Table 1. Experimental Design Summary

Parameter	Value(s)
Null Hypothesis Prevalence (p_0)	25%, 50%, 75%
Test Sample Sizes	50, 75, 100
Number of Simulations	2000
Tests	EM LRT, EM Bootstrap CI, t-test, KS-test
Metrics	AUC, FPR

4 Results

This section presents the results of the experiments comparing different hypothesis testing methods across various datasets and calibration settings. We first present the classification performance and training set prevalence on the

datasets to provide context for the hypothesis testing results. We then report the AUC and FPR by dataset, test sample size, calibration, and unlabled prevalence hypothesis testing method.

Table 2. Classification performance and training set class prevalence

Dataset	AUC	Train Prevalence	Train Set Size (N)
Diabetes	0.827	34.8%	385
Cancer	0.993	37.2%	285
Simulated	0.726	50.0%	1000

Table 2 demonstrates that the underlying classifier has a breadth of performance across datasets. Note that, despite evaluating both a calibrated and uncalibrated logistic regression models, we only report a single AUC. This is because Platt's calibration method does not change the AUC of the underlying classifier. The Breast Cancer Wisconsin (Cancer) dataset can be almost perfectly classified using a logistic regression model. Classifier performance was worse for the Pima Indians Diabetes (Diabetes) dataset, and worse still for the simulated dataset. Prevalence of the positive class was similar in the two real-world datasets (35%), while the simulated dataset had a balanced training set. There is a range of training set sizes across the datasets, with the simulated dataset being over three times the size of the Diabetes dataset.

Table 3. Area under the receiver operating curve and false positive rate for the Pima Indian Diabetes dataset. Top performing methods are in **bold.**

		AUC				False Positive Rate			
Calibration	test-size	EM LRT	Bootstrap CI	t-test	KS -test	EM LRT	Bootstrap CI	t-test	KS -test
Platt's	50	**0.6034**	0.6019	0.5830	0.5770	**6.25%**	7.38%	6.48%	8.45%
	75	**0.6368**	0.6351	0.6105	0.6039	**7.63%**	8.28%	7.83%	10.73%
	100	**0.6623**	0.6604	0.6359	0.6195	**7.75%**	8.48%	8.23%	11.63%
None	50	**0.6283**	0.6238	0.5869	0.5756	9.00%	11.55%	**6.43%**	8.50%
	75	**0.6577**	0.6528	0.6143	0.5997	11.23%	12.90%	**7.70%**	11.25%
	100	**0.6847**	0.6790	0.6410	0.6152	13.65%	14.93%	**8.07%**	12.28%

Table 3 displays the AUC and FPR for the Pima Indian Diabetes dataset. This dataset had a moderate number of samples (385 in the training set), and the underlying classifier performance was moderate (AUC: 0.827). For this dataset, EM LRT was the most skilled (AUC: 0.6034 - 0.6847), followed by the EM Bootstrap CI (AUC: 0.6019 - 0.6790) across different test sizes. The t-test was marginally inferior (AUC: 0.5830 - 0.6410), and the KS-test was substantially

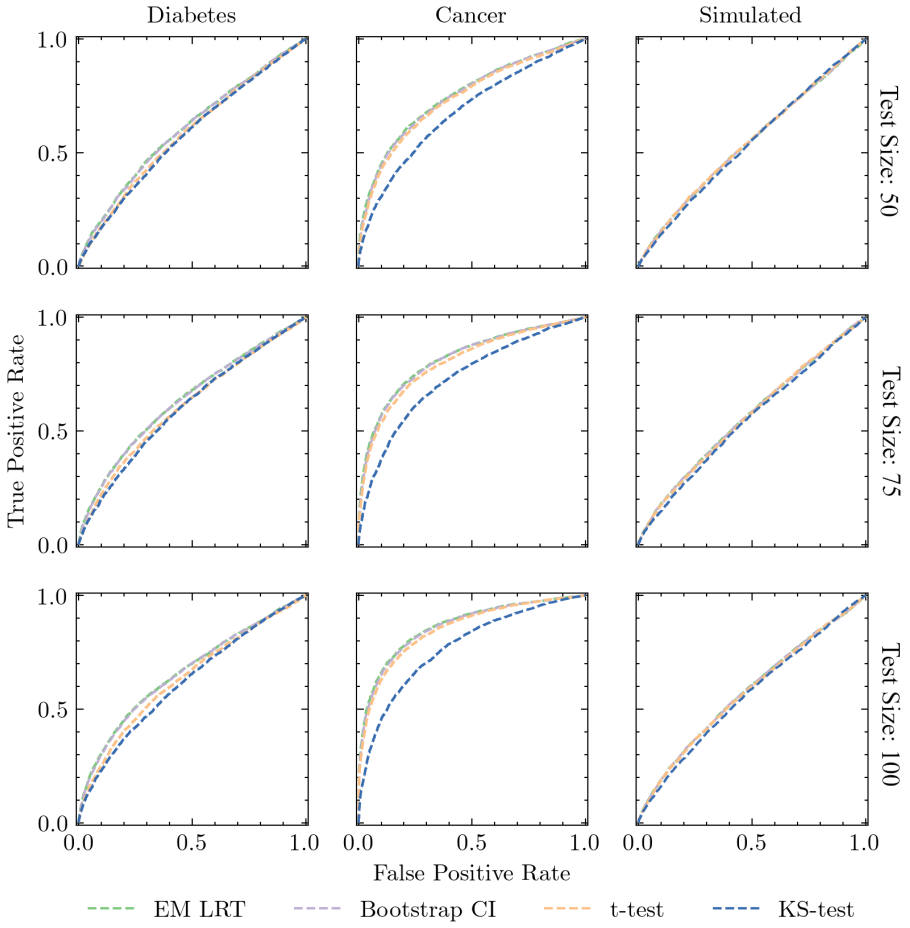

Fig. 1. Receiver operating characteristic curves show that across all tests, skill improves as sample size increases and performance of the underlying classifier improves.

worse (0.5770 - 0.6152). Notably, the FPRs for the EM LRT, EM Bootstrap CI, and t-test with calibration were close to or slightly above the target 5%, ranging from 6.25% to 8.48%, while the KS-test was poorly calibrated. Without calibration, EM LRT and EM Bootstrap CI achieved much higher FPRs (9.00% - 14.93%).

Table 4. Area under the receiver operating characteristic curve and false positive rate for the Breast Cancer Wisconsin dataset. Top performing methods are in **bold.**

		AUC				False Positive Rate			
Calibration	test-size	EM LRT	Bootstrap CI	t-test	KS -test	EM LRT	Bootstrap CI	t-test	KS -test
Platt's	50	**0.7525**	0.7491	0.7387	0.6781	**5.20%**	6.07%	5.67%	8.47%
	75	**0.8193**	0.8165	0.8036	0.7314	**5.17%**	6.08%	5.30%	10.90%
	100	**0.8667**	0.8633	0.8527	0.7727	**5.65%**	6.43%	5.88%	13.20%
None	50	**0.7555**	0.7462	0.7284	0.6795	7.23%	9.28%	**5.78%**	8.33%
	75	**0.8129**	0.8040	0.7899	0.7316	7.95%	9.80%	**5.67%**	10.85%
	100	**0.8478**	0.8382	0.8411	0.7751	9.65%	11.23%	**6.47%**	13.00%

Table 4 shows the AUC and FPR for the Breast Cancer Wisconsin dataset. This dataset had a moderate number of samples (285 in the training set), and the underlying classifier performance was nearly perfect (AUC: 0.993). The EM LRT method achieved the highest AUC across all evaluated conditions, indicating superior discriminatory power. This superiority was maintained regardless of test size or the application of Platt's calibration. As the test size increased from 50 to 100, the AUC for all methods showed a corresponding improvement. For example, with Platt's calibration, the AUC for the EM LRT increased from 0.7525 to 0.8667. Analysis of the false positive rates (FPR) revealed that the optimal method was dependent on the calibration strategy. When Platt's scaling was applied, the EM LRT method consistently produced the lowest FPR, with rates as low as 5.17%. In contrast, without calibration, the t-test yielded the lowest FPR, maintaining a rate between 5.67% and 6.47%. In all scenarios, the KS-test exhibited the highest FPR.

Table 5 presents the AUC and FPR for the simulated dataset. This dataset had a large number of samples (1000 in the training set) and the underlying classifier performance was poor (AUC: 0.726). In contrast to the other datasets, no single method consistently achieved the highest AUC. Top performing methods in terms of discriminatory power varied with test size. For example, under Platt's calibration, the t-test was marginally superior at a test size of 50 (AUC = 0.5460), while the EM LRT and EM Bootstrap CI methods performed best at test sizes of 75 (AUC = 0.5651) and 100 (AUC = 0.5781), respectively. A consistent trend across all methods was the modest improvement in AUC with increasing test size. The KS-test consistently yielded the lowest AUC values.

Table 5. Area under the receiver operating characteristic curve and false positive rate for the simulated dataset. Top performing methods are in **bold.**

		AUC				False Positive Rate			
Cali-bration	test-size	EM LRT	Bootstrap CI	t-test	KS -test	EM LRT	Bootstrap CI	t-test	KS -test
Platt's	50	0.5445	0.5446	**0.5460**	0.5390	5.70%	7.13%	5.95%	6.18%
	75	**0.5651**	0.5648	0.5622	0.5484	5.95%	6.62%	5.87%	6.75%
	100	0.5779	**0.5781**	0.5753	0.5628	6.38%	6.92%	6.52%	7.45%
None	50	0.5449	0.5439	**0.5457**	0.5388	**5.20%**	8.08%	5.95%	6.37%
	75	**0.5636**	0.5625	0.5627	0.5487	6.07%	7.50%	**6.02%**	7.07%
	100	0.5742	0.5730	**0.5744**	0.5619	7.23%	8.20%	**6.45%**	7.58%

The control of the FPR was also variable. Without calibration, the EM LRT was most effective at the smallest test size (FPR = 5.20%), whereas the t-test was most effective at larger test sizes. With Platt's scaling, no single method consistently produced the lowest FPR. Overall, on the simulated dataset, the EM LRT, Bootstrap CI, and t-test methods yielded comparable results, with the optimal choice for both AUC and FPR being contingent on the specific test size.

5 Discussion

When using the outputs of a well calibrated logistic regression model, the EM LRT method most consistently yielded a false positive rate close to the specified 5%, though it was consistently above 5%. The EM Bootstrap CI method consistently resulted in a false positive rate higher than the specified 5%. This is in disagreement with Tasche's findings, which generally found that bootstrapping the EM estimate had adequate coverage [22]. This disagreement might arise from the fact that we are evaluating the confidence interval at a larger width (95%) than Tasche (90%), leading to less stable estimates. Additionally, Tasche only simulated 100 runs compared to our 2000, and it's possible that the minor FPR excess (no more than 3% among calibrated models) did not appear in Tasche's study due to stochasticity.

Calibration proved to be important in achieving false positive rate close to the expected value for the methods involving EM, particularly among the real-world datasets with smaller sample sizes. This is in line with Esuli et al., who found performance of point estimates of prevalence using the EM algorithm were highly sensitive to calibration of underlying classifiers [10]. The t-test method performance was robust to calibration status, performing nearly identically when applied to the calibrated vs uncalibrated model outputs. In applications where reliable calibration of the underlying classifier is not tenable, a t-test could be utilized.

The EM LRT demonstrated the greatest skill in terms of area under the receiver operating characteristic curve across the two benchmark datasets and all sample sizes. Detection skill was correlated with the skill of the underlying classifier, and the sample size, consistent with previous work [18, 22]. While the EM LRT performed the best in our experiments, it can only be utilized in two-sided hypothesis testing. For one-sided hypothesis testing, the EM Bootstrap CI could be employed as it performed similarly to EM LRT.

There are several limitations to this study. Use of an artificial prevalence protocol may create unreasonable and highly unlikely prevalences that do not reflect reality [12]. Additionally, we have not considered types of dataset shift outside of prior shift. Covariate shift and concept shift would likely affect our results, as González et al. found the EM algorithm to degrade in performance in the presence of these shifts [15]. Finally, our analysis was limited to binary classes and neglected more complex topics in quantification including multiclass quantification (for example, see [4]) or ordinal quantification (for example see [5]).

6 Conclusion

This paper systematically compared various methods for unlabeled class prevalence hypothesis testing in binary classification tasks using both real-world and simulated datasets. The findings highlight the importance of calibration, particularly for EM based methods, in achieving false positive rates that align with the desired 5% target. EM LRT generally provided the greatest discriminative power with the best calibration. While the EM LRT consistently produced false positive rates close to 5% with a well-calibrated logistic regression model, the EM Bootstrap Confidence Interval (EM Bootstrap CI) method often resulted in higher FPRs, possibly due to the wider confidence interval evaluated and increased simulation runs compared to prior work. The t-test, in contrast, demonstrated robust performance independent of calibration status. This empirical evaluation provides valuable insights into the performance characteristics of different prior shift detection techniques, contributing to the limited existing literature on this crucial task.

References

1. Breast Cancer Wisconsin (Diagnostic) Data Set, `https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data`
2. Pima Indians Diabetes Database, `https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database`
3. Alexandari, A., Kundaje, A., Shrikumar, A.: Maximum Likelihood with Bias-Corrected Calibration is Hard-To-Beat at Label Shift Adaptation (Jun 2020). `https://doi.org/10.48550/arXiv.1901.06852`, `http://arxiv.org/abs/1901.06852`, arXiv:1901.06852 [cs]

4. Bunse, M.: On multi-class extensions of adjusted classify and count. In: Proceedings of the 2nd International Workshop on Learning to Quantify (LQ 2022). pp. 43–50 (2022), https://lq-2022.github.io/proceedings/Bunse2022b.pdf
5. Bunse, M., Moreo, A., Sebastiani, F., Senz, M.: Ordinal Quantification Through Regularization. In: Amini, M.R., Canu, S., Fischer, A., Guns, T., Kralj Novak, P., Tsoumakas, G. (eds.) Machine Learning and Knowledge Discovery in Databases. pp. 36–52. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-26419-1_3
6. Castaño, A., Alonso, J., González, P., Coz, J.J.d.: An Equivalence Analysis of Binary Quantification Methods. Proceedings of the AAAI Conference on Artificial Intelligence **37**(6), 6944–6952 (Jun 2023). https://doi.org/10.1609/aaai.v37i6.25849, https://ojs.aaai.org/index.php/AAAI/article/view/25849
7. Daughton, A.R., Paul, M.J.: A bootstrapping approach to social media quantification. Social Network Analysis and Mining **11**(1), 73 (Aug 2021). https://doi.org/10.1007/s13278-021-00760-0, https://doi.org/10.1007/s13278-021-00760-0
8. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum Likelihood from Incomplete Data Via the *EM* Algorithm. Journal of the Royal Statistical Society Series B: Statistical Methodology **39**(1), 1–22 (Sep 1977). https://doi.org/10.1111/j.2517-6161.1977.tb01600.x, https://academic.oup.com/jrsssb/article/39/1/1/7027539
9. Esuli, A., Fabris, A., Moreo, A., Sebastiani, F.: Learning to Quantify, The Information Retrieval Series, vol. 47. Springer International Publishing, Cham (2023). https://doi.org/10.1007/978-3-031-20467-8, https://link.springer.com/10.1007/978-3-031-20467-8
10. Esuli, A., Molinari, A., Sebastiani, F.: A Critical Reassessment of the Saerens-Latinne-Decaestecker Algorithm for Posterior Probability Adjustment. ACM Transactions on Information Systems **39**(2), 1–34 (Apr 2021). https://doi.org/10.1145/3433164, https://dl.acm.org/doi/10.1145/3433164
11. Esuli, A., Moreo, A., Sebastiani, F., Sperduti, G.: A detailed overview of LeQua 2022: learning to quantify. In: Working Notes of the 13th conference and labs of the evaluation forum (CLEF 2022). Bologna, Italy (2022)
12. Esuli, A., Sebastiani, F.: Optimizing Text Quantifiers for Multivariate Loss Functions. ACM Trans. Knowl. Discov. Data **9**(4), 27:1–27:27 (Jun 2015). https://doi.org/10.1145/2700406, https://doi.org/10.1145/2700406
13. Evans, C., G'Sell, M.: Sequential label shift detection in classification data: An application to dengue fever. PLOS ONE **19**(9), e0310194 (Sep 2024). https://doi.org/10.1371/journal.pone.0310194, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11404796/
14. Gart, J.J., Buck, A.A.: Comparison of a screening test and a reference test in epidemiologic studies. II. A probabilistic model for the comparison of diagnostic tests. American Journal of Epidemiology **83**(3), 593–602 (May 1966). https://doi.org/10.1093/oxfordjournals.aje.a120610
15. González, P., Moreo, A., Sebastiani, F.: Binary quantification and dataset shift: an experimental investigation. Data Mining and Knowledge Discovery **38**(4), 1670–1712 (Jul 2024). https://doi.org/10.1007/s10618-024-01014-1, https://doi.org/10.1007/s10618-024-01014-1
16. Hopkins, D.J., King, G.: A Method of Automated Nonparametric Content Analysis for Social Science. American Journal of Political Science **54**(1), 229–247 (2010). https://doi.org/10.1111/j.1540-5907.2009.00428.x,

https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-5907.2009.00428.x, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-5907.2009.00428.x

17. Lipton, Z., Wang, Y.X., Smola, A.: Detecting and Correcting for Label Shift with Black Box Predictors. In: Proceedings of the 35th International Conference on Machine Learning. pp. 3122–3130. PMLR (Jul 2018), https://proceedings.mlr.press/v80/lipton18a.html, iSSN: 2640-3498
18. Maletzke, A., Hassan, W., Reis, D.d., Batista, G.: The Importance of the Test Set Size in Quantification Assessment. vol. 3, pp. 2640–2646 (Jul 2020). https://doi.org/10.24963/ijcai.2020/366, https://www.ijcai.org/proceedings/2020/366, iSSN: 1045-0823
19. Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers pp. 61–74 (1999)
20. Rabanser, S., Günnemann, S., Lipton, Z.: Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift. In: Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019), https://papers.nips.cc/paper_files/paper/2019/hash/846c260d715e5b854ffad5f70a516c88-Abstract.html
21. Saerens, M., Latinne, P., Decaestecker, C.: Adjusting the Outputs of a Classifier to New a *Priori* Probabilities: A Simple Procedure. Neural Computation **14**(1), 21–41 (Jan 2002). https://doi.org/10.1162/089976602753284446, https://direct.mit.edu/neco/article/14/1/21-41/6577
22. Tasche, D.: Confidence Intervals for Class Prevalences under Prior Probability Shift. Machine Learning and Knowledge Extraction **1**(3), 805–831 (Sep 2019). https://doi.org/10.3390/make1030047, https://www.mdpi.com/2504-4990/1/3/47, number: 3 Publisher: Multidisciplinary Digital Publishing Institute