A new generic method to improve machine learning applications in official statistics

Mr. Kevin Kloos
Intern Methodologist
Statistics Netherlands (CBS)
29 August 1997 (23 years)
https://www.cbs.nl/en-gb/
the Hague, the Netherlands
kevinkloos@hotmail.com

Abstract

The use of machine learning algorithms at national statistical institutes has increased significantly over the past few years. Applications range from new imputation schemes to new statistical output based entirely on machine learning. The results are often very promising. However, recent studies have shown that the use of machine learning in official statistics always introduces a bias, known as misclassification bias. That type of bias does not occur in traditional applications of machine learning and therefore it has received little attention in the academic literature. In earlier work, we have collected existing methods that are able to correct misclassification bias. We have compared their statistical properties, including bias, variance and mean squared error. In this paper, we present a new generic method to correct misclassification bias and we derive its statistical properties. Moreover, we show numerically that it has a lower mean squared error than the existing alternatives in a wide variety of settings. We believe that our new method may improve machine learning applications in official statistics and we aspire that our work will stimulate further methodological research in this area.

Submission for the 2021 IAOS Prize for Young Statisticians

1 Introduction

National statistical institutes currently apply many different types of machine learning algorithms. Classification algorithms are the most popular type. Examples of classical classification algorithms are logistic regression and linear discriminant analysis, but new innovative algorithms have been introduced over the last decades, like additive models, decision trees and deep learning (Friedman et al., 2001). The aim of classification algorithms is to minimize a chosen loss function. However, classifying objects individually can lead to biased results when generalizing these individual objects to aggregate statistics, like a proportion of the population (Schwarz, 1985; Scholtus and van Delden, 2020). The accuracy of an aggregate statistic depends, among other things, on the classification probabilities and the distribution of objects over the classes. It is easy to imagine that a class with a low number of objects can make relatively equal errors as a class with a large number of objects. While the relative error can be the same between both classes, the difference in absolute numbers can be much larger. The imbalance of the classification errors can lead to misclassification bias. As a first step, we focus on obtaining an accurate estimate of the proportion of objects that belong to our class of interest, i.e., the base rate.

We can reduce the misclassification bias of the base rate with correction methods. In a previous paper, we compared the statistical properties of these five correction methods and showed that the calibration estimator minimizes mean squared error (Kloos et al., 2021). However, the result from that paper does not generalize to time series, because the base rate will change over time. The target populations that are interesting for national statistical institutes, where we produce statistics on a monthly, quarterly or annual basis, change over time. Allowing the base rate to change over time while still imposing the assumption that the classification probabilities are constant over time, is a special case of concept drift called prior probability shift (Moreno-Torres et al., 2012).

The most effective, and most costly and time-consuming, solution to deal with prior probability shift, is to generate a completely new test set every time we produce a new base rate. A more efficient solution is to construct a test set and recycle it over time. In contrast to generating new test sets in each time period, we cannot assume that a test set is a simple random sample of the target population if the base rate changes over time. Therefore, new expressions for the bias and variance are needed to evaluate the MSE of the five correction methods. These expressions were previously computed by Meertens (2020). The main contribution of this paper is a new generic method to correct for misclassification bias when dealing with prior probability shift. We will refer to the resulting estimator as *the mixed estimator*, because it combines the strengths of two estimators based on existing correction methods. We will derive (approximate) closed-form expressions for the bias and variance of the mixed estimator. Moreover, we will numerically compare the mean squared error of the mixed estimator with existing methods.

The remainder of the paper is organized as follows: In Chapter 2, we introduce the problem and assumptions and we recap the properties of the original correction methods. Chapter 3 introduces the mixed estimator. Moreover, we will compare the mixed estimator with the original correction methods. Chapter 4 contains a discussion and conclusion of this paper.

2 Model under prior-probability shift

In this paper, we use of the same mathematical assumptions as in Kloos et al. (2021). Before we dive into the mathematical expressions, we briefly discuss the terminology of the upcoming sections. Our parameter of interest is the proportion of objects that belong to class 1 in the target population, denoted as α. The target population has N objects, which are belong to one of two classes, either class 0 or class 1. Similarly to Van Delden et al. (2016), we assume that the classifier has a probability of p_{00} to correctly classify an object of class 0 and a probability of p_{11} to correctly label an object of class 1. These probabilities are unknown in practice, so we randomly sample a test set of size n from the target population where both the true labels and the estimated labels are known. We assume that the size of the test set is much smaller than the target population, such that we can assume that sampling the test set from the target population follows a binomial distribution, instead of a hyper-geometric distribution. Further details are provided by Chapter 2 of Kloos et al. (2021), but are not needed to understand the concepts of this paper.

Kloos et al. (2021) assumes that the prior probability shift is zero. Here, we allow for a nonzero prior. We introduce the following notation. First, we need to distinguish a target population U at time 0 with a target population U' at time t. Therefore, we can define α' as the base rate of the target population U'. Moreover, the classification probabilities of the target population are equal for U and U', so the base rate α' is the only new parameter in this paper. Therefore, we need to define α more sharply to the true base rate of the target population U, whereof the test set is constructed.

Before we describe the differences between Kloos et al. (2021) and Meertens (2020), we briefly introduces the correction methods. The baseline estimator computes the proportion of objects in the test set that belong to class 1. The classify-and-count estimates computes the proportion of objects that are classified by the machine learning algorithm to class 1 in the target population. The subtracted-bias estimator estimates the bias by estimating the classification probabilities in the test set and the classify-and-count estimator. We compute the subtracted-bias estimator by subtracting this estimated bias from the classify-and-count estimator. The misclassification estimator multiplies the inverted row-normalised test set by the classify-and-count estimator. Last, the calibration estimator multiplies the column-normalised test set by the classify-and-count estimator.

The closed-form expressions in Meertens (2020) show that prior probability shift

does not affect the RMSE of the classify-and-count estimator and subtracted-bias estimator majorly. The classify-and-count estimator does not use the test set, so a different base rate does not affect this estimator. The subtracted-bias estimator only uses the test set to estimate the classification probabilities and therefore the estimator is not affected largely under prior probability shift. However, the baseline estimator and the calibration estimator are biased under prior probability shift. This is in contrast to the situation under a fixed base rate, where the baseline estimator and the calibration estimator are unbiased. Therefore, the misclassification estimator remains the only estimator that is (asymptotically) unbiased. According to Kloos et al. (2021) and Meertens (2020), the misclassification estimator has a high variance when the classification probabilities are low, but the variance changes slightly under prior probability shift. All in all, none of these correction methods have a consistently low RMSE. In the next chapter, we will introduce a new estimator that performs better than the five original correction methods.

3 Mixed estimator

In this section, we introduce a new estimator: the mixed estimator. The mixed estimator is a combination between the misclassification estimator (Buonaccorsi, 2010) and the calibration estimator (Kuha and Skinner, 1997). In Kloos et al. (2021) and Meertens (2020), we denoted that the calibration estimator is unbiased under a fixed base rate, but biased under prior probability shift. The misclassification estimator has a higher variance, but the RMSE remains fairly stable under prior probability shift. These two properties can be combined: as initial starting point, we take the calibration estimator α_c at time 0, but we add the difference between the misclassification estimator at time t (α'_p) and time 0 (α_p). Therefore, the expression for the mixed estimator is:

$$\begin{aligned}\hat{\alpha}'_m &= \hat{\alpha}_c + \left[\hat{\alpha}'_p - \hat{\alpha}_p\right] \\ &= \frac{n_{10}}{n_{+0}}(1-\hat{\alpha}^*) + \frac{n_{11}}{n_{+1}}\hat{\alpha}^* + \frac{(\hat{\alpha}')^* - \hat{\alpha}^*}{\hat{p}_{00}+\hat{p}_{11}-1}\end{aligned} \tag{1}$$

To the best of our knowledge, this is the first paper where the mixed estimator is introduced. Therefore, the closed-form expressions for bias and variance that we have derived are new as well.

Theorem 1. *The mixed estimator $\hat{\alpha}'_m$ is a biased, but consistent, estimator for $\alpha' \neq \alpha$:*

$$B\left[\hat{\alpha}'_m\right] = \frac{(\alpha'-\alpha)(V(\hat{p}_{00})+V(\hat{p}_{11}))}{(p_{00}+p_{11}-1)^2} + O\left(\frac{1}{n^2}\right). \tag{2}$$

The variance of $\hat{\alpha}'_m$ is equal to the following expression, where the expressions of $V(\hat{p}_{00})$

and $V(\hat{p}_{11})$ can be found in the first pages of the Appendix:

$$
\begin{aligned}
V(\hat{\alpha}'_m) = &\frac{\alpha p_{11}}{n} \times \left(1 - \frac{\alpha p_{11}}{(1-\alpha)(1-p_{00}) + \alpha p_{11}}\right) \\
&+ \frac{(1-\alpha)p_{00}}{n} \times \left(1 - \frac{(1-\alpha)p_{00}}{(1-\alpha)p_{00} + \alpha(1-p_{11})}\right) \\
&+ (\alpha' - \alpha)^2 \times \frac{V(\hat{p}_{00}) + V(\hat{p}_{11})}{(p_{00} + p_{11} - 1)^2} \\
&+ (\alpha' - \alpha) \times \left[\frac{\alpha p_{00}(1-p_{00})(1-p_{11}) + (1-\alpha)p_{00}p_{11}(1-p_{11})}{n(p_{00} - \alpha(p_{00} + p_{11} - 1))(p_{00} + p_{11} - 1)}\right. \\
&\quad \left. - \frac{\alpha p_{00}(1-p_{00})p_{11} + (1-\alpha)(1-p_{00})p_{11}(1-p_{11})}{n((1-\alpha)(1-p_{00}) + \alpha p_{11})(p_{00} + p_{11} - 1)}\right] + O\left(\frac{1}{n^2}\right). \quad (3)
\end{aligned}
$$

Proof. See the online Appendix on `https://www.github.com/kevinkloos/IAOS`. □

From Theorem 1, we see that the mixed estimator is slightly biased, but consistent. The variance function is complex, but we can see that the variance will be bigger when the difference between α' and α. To get a better overview of this mixed estimator, we will perform three simulation studies.

In the first simulation study, we consider a class-balanced dataset ($\alpha = 0.5$), with a small test set of size $n = 1000$, a large population dataset of $N = 3 \times 10^5$ and a rather poor classifier having classification probabilities $p_{00} = 0.6$ and $p_{11} = 0.7$. From Figure 1, we can see that the mixed estimator is in general a stable estimator with a low amount of bias and much less variance than the misclassification estimator. However, the variance of the mixed estimator tends to increase when the difference between α' and α gets larger, which is in line with the observations in the previous section. The mixed estimator performs much better than the misclassification estimator and the calibration estimator: it has almost no bias and has much less variance than the misclassification estimator.

A situation where the mixed estimator does not work as well as expected, can be found in Figure 2. We specify the following parameters: $p_{00} = 0.94$, $p_{11} = 0.97$, $\alpha = 0.98$, $n = 1000$ and $N = 3 \times 10^5$. The misclassification estimator tends to have more extreme outliers when the difference between α' and α increases. This affects the mixed estimator in terms of variance. Furthermore, the mixed estimator can predict values outside the $[0, 1]-$interval. Obviously, we cannot encounter these values in practice and it is therefore a problem that we obtain these estimates. It seems that this problem occurs less often for the mixed estimator than for the misclassification estimator and the mixed estimator performs still better than the misclassification estimator and the calibration estimator. Finally, we can observe that the variance of the mixed estimator is always

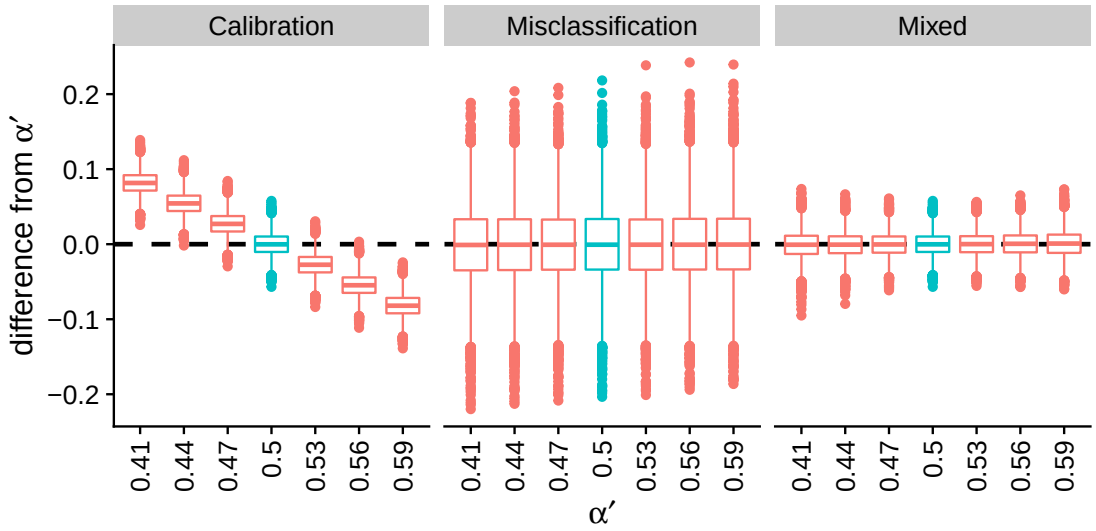

Figure 1: Simulation study to observe the change in prediction error under concept drift using boxplots. The calibration, misclassification and mixed estimator are compared given an initial base rate $\alpha = 0.5$ (blue) and different values of α' (red). The x-axis shows the different base rates and the y-axis shows the distribution of the difference from α'. All the parameters: $p_{00} = 0.6$, $p_{11} = 0.7$, $n = 1000$ and $N = 3 \times 10^5$.

lower than the variance of the misclassification estimator. Despite the outliers, it is still the best estimator out of the three.

In the first two simulation studies, the misclassification estimator did not work properly, and we showed values of α' that are close to α. It is also interesting to see what happens when the misclassification estimator has a low RMSE for α and what happens when α' differs substantially from α. We perform a simulation study with $\alpha = 0.75$, $p_{00} = 0.85$, $p_{11} = 0.90$, $n = 1000$ and $N = 3 \times 10^5$, shown in Figure 3. We observe that the distribution of the mixed estimator is similar to the distribution of the misclassification estimator. The reason behind is, is that the misclassification estimator performs similarly to the calibration estimator at time zero. However, the figures and the numbers show that the mixed estimator still performs consistently better than the misclassification estimator.

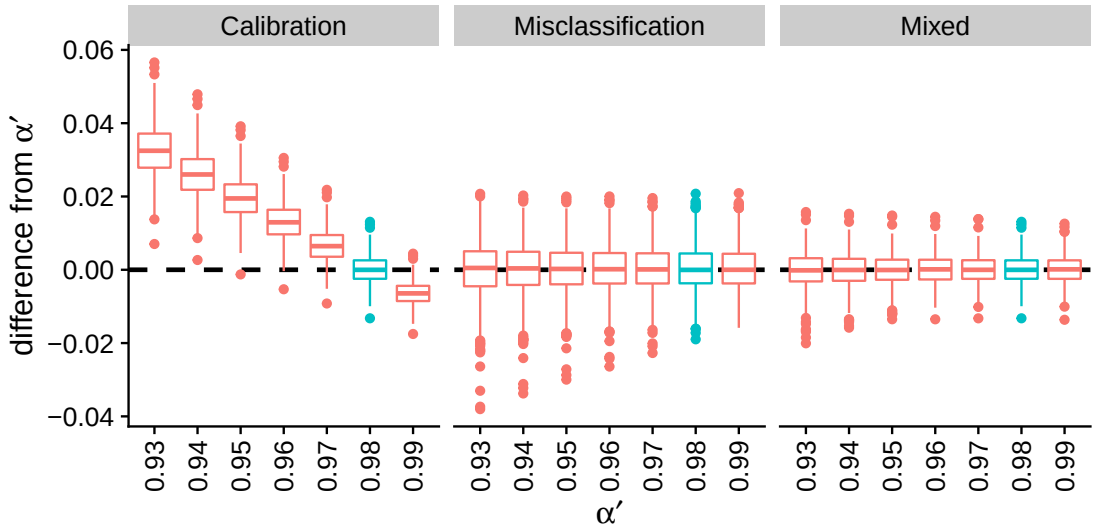

Figure 2: Simulation study to observe the change in prediction error under concept drift using boxplots. The calibration, misclassification and mixed estimator are compared given an initial base rate $\alpha = 0.98$ (blue) and different values of α' (red). The x-axis shows the different base rates and the y-axis shows the distribution of the difference from α'. All the parameters: $p_{00} = 0.94$, $p_{11} = 0.97$, $n = 1000$ and $N = 3 \times 10^5$.

4 Conclusion and discussion

Thus, we may conclude that our mixed estimator outperforms the estimators currently available in the academic literature. The mixed estimator has less bias than the calibration estimator and less variance than the calibration estimator. The mixed estimator performs much better than the original correction methods when classification probabilities of a classification algorithm are not extremely low and when the variance of the misclassification estimator is much larger than the variance of the calibration estimator. Our results show that the mixed estimator outperforms both the calibration estimator and the misclassification estimator in any dataset and for any classification algorithm used.

Even though that the new mixed estimator performs better than the original correction methods, we still believe that the correction methods might be improved further. We could construct a new estimator by combining biased, but invariant correction methods. New research directions lay in combining the correction methods in such a way that both bias and variance of the new estimator will be consistently low.

The estimator could be extended for correction methods than can predict more than two classes. The downside is that the amount of parameters increases quadratically and the quality measure should be adapted for multiple classes. A possible solution is to

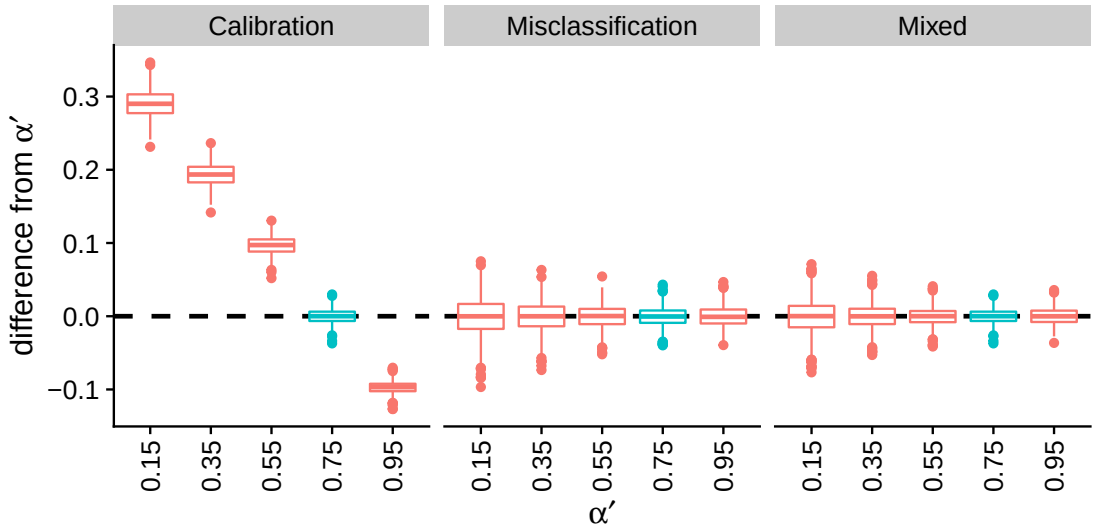

Figure 3: Simulation study to observe the change in prediction error under concept drift using boxplots. The calibration, misclassification and mixed estimator are compared given an initial base rate $\alpha = 0.75$ (blue) and different values of α' (red). The x-axis shows the different base rates and the y-axis shows the distribution of the difference from α'. All the parameters: $p_{00} = 0.85$, $p_{11} = 0.90$, $n = 1000$ and $N = 3 \times 10^5$.

further elaborate the simulation studies, instead of computing closed-form mathematical expressions. A final extension that we recommend is to allow the classification probabilities to differ between the objects within a group.

With this paper, we hope to have raised awareness that aggregating outcomes of machine learning algorithms can be very inaccurate, even if the algorithms has a high prediction accuracy. Furthermore, this paper is an addition to scientific literature on theory of misclassification bias. Finally, we proposed a new generic method that can be used by NSIs to improve machine learning applications within official statistics.

References

Buonaccorsi, J. P. (2010). *Measurement Error: Models, Methods, and Applications*. Chapman & Hall/CRC, Boca Raton, FL.

Friedman, J. H., Hastie, T., Tibshirani, R., et al. (2001). *The elements of statistical learning*, volume 1. Springer, New York.

Kloos, K., Meertens, Q. A., Scholtus, S., and Karch, J. D. (2021). Comparing correction methods to reduce misclassification bias.

Knottnerus, P. (2003). *Sample Survey Theory: Some Pythagorean Perspectives*. Springer Science & Business Media.

Kuha, J. and Skinner, C. J. (1997). Categorical Data Analysis and Misclassification. In Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., and Trewin, D., editors, *Survey Measurement and Process Quality*, pages 633–670. Wiley.

Meertens, Q. A. (2020). Understanding the output quality of official statistics that are based on machine learning algorithms.

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., and Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530.

Scholtus, S. and van Delden, A. (2020). On the Accuracy of Estimators Based on a Binary Classifier. Discussion Paper, Statistics Netherlands, The Hague.

Schwarz, J. E. (1985). The Neglected Problem of Measurement Error in Categorical Data:. *Sociological Methods & Research*.

Van Delden, A., Scholtus, S., and Burger, J. (2016). Accuracy of Mixed-Source Statistics as Affected by Classification Errors. *Journal of Official Statistics*, 32(3):619–642.