Minimising quantifier variance under prior probability shift

Dirk Tasche[1]

[1]Independent Researcher, Zurich, Switzerland

Abstract

For the binary prevalence quantification problem under prior probability shift, we determine the asymptotic variance of the maximum likelihood estimator. We find that it is a function of the Brier score for the regression of the class label on the features under the test data set distribution. This observation suggests that optimising the accuracy of a base classifier, as measured by the Brier score, on the training data set helps to reduce the variance of the related quantifier on the test data set. Therefore, we also point out training criteria for the base classifier that imply optimisation of both of the Brier scores on the training and the test data sets.

Keywords

Prior probability shift, quantifier, class distribution estimation, Cramér-Rao bound, maximum likelihood estimator, Brier score

1. Introduction

The survey paper [1] described the problem to estimate prior class probabilities (also called prevalences) on a test set with a different distribution than the training set (the *quantification* problem) as "Given a labelled training set, induce a quantifier that takes an unlabelled test set as input and returns its best estimate of the class distribution." As becomes clear from [1] and also more recent work on the problem, it has been widely investigated in the past twenty years.

A lot of different approaches to quantification of prior class probabilities has been proposed and analysed (see, e.g. [1, 2, 3]), but it appears that the following question has not yet received very much attention:

Is it worth the effort to try to train a good (accurate) hard (or soft or probabilistic) classifier as the 'base classifier' for the task of quantification if the class labels of individual instances are unimportant and only the aggregate prior class probabilities are of interest?

In principle, there is a clear answer to this question. The accuracy of the classifier matters at least in the extreme cases:

- If a classifier is least accurate because its predictions and the true class labels are stochastically independent, then quantification is not feasible.
- If a classifier is most accurate in the sense of making perfect predictions then perfect quantification is easy by applying Classify & Count [4].

But if no perfect classifier is around, can we be happy to deploy a moderately accurate classifier for quantification or should we rather strive to develop an optimal classifier, possibly based on an comprehensive feature selection process?

Some researchers indeed suggest that the accuracy of the base classifiers is less important for quantification than for classification. [5] made the following statements:

- From the abstract of [5]: "These strengths can make quantification practical for business use, even where classification accuracy is poor."
- P. 165 of [5]: "The effort to develop special purpose features or classifiers could increase the cost significantly, with no guarantee of an accurate classifier. Thus, an imperfect classifier is often all that is available."
- P. 166 of [5]: "It is sufficient but not necessary to have a perfect classifier in order to estimate the class distribution well. If the number of false positives balances against false negatives, then the overall count of predicted positives is nonetheless correct. Intuitively, the estimation task is easier for not having to deliver accurate predictions on individual cases."

The point on the mutual cancellation of false positives and false negatives is mentioned also by a number of other researchers like for instance [6]. On p. 74, [6] wrote: "Equation 1 [with the definition of the F-measure] shows that F_1 deteriorates with $(FP + FN)$ and not with $|FP - FN|$, as would instead be required of a function that truly optimizes quantification."

There are also researchers that hold the contrary position, at least as quantification under an assumption of prior probability shift (see (1) below for the formal definition) is concerned:

- [7] noted for the class of 'ratio estimators' they introducted that it was both desirable and feasible to construct estimators with small asymptotic variances.

Learning to Quantify: Methods and Applications (LQ 2021)
dirk.tasche@gmx.net (D. Tasche)
0000-0002-2750-2970 (D. Tasche)

- [8] demonstrated by a simulation study that estimating the prior class probabilities by means of a more accurate base classifier may entail much shorter confidence intervals for the estimates.
- [9, 10] pointed out the efficiency of the maximum likelihood estimator (MLE) for the quantification task and made cases for its application.

In the following, we revisit the question of the usefulness of accurate classifiers for quantification:

After giving an overview of related research in Section 2 and specifying the setting and assumptions for the binary quantification problem in Section 3, we recall the technical details of the definition of the MLE for the positive class prevalence in Section 4. In particular, we show that the MLE is well-defined under the mild condition that the test sample consists of at least two different points, see (6) below.

In Section 5, we describe, based on its representation in terms of the Fisher information, the Cramér-Rao lower bound for the variances of unbiased estimators of the positive class prevalence, see (9) below. This lower bound, at the same time, is the large sample variance of the MLE defined in Section 4. Thus, the Cramér-Rao lower bound is achievable in theory by MLEs – which might explain to some extent the superiority of MLEs as observed in [9, 10].

In Section 6, we show for the test distribution of the feature vector that its associated Fisher information with respect to the positive class prevalence is – up to a factor depending only on the true prevalence – just the variance of the posterior positive class probability (Proposition 6.1 below). The variance of the posterior probability is closely related to the Brier score for the regression of the class label on the feature vector. While the Brier score decreases when the information content of the feature vector increases, the variance of the posterior probability decreases with shrinking information content of the features. In any case, these observations imply that the large sample variance of the MLE (or the Cramér-Rao lower bound for the variances of unbiased estimators) may be reduced when a feature vector with larger information content is selected. Section 6 concludes with two suggestions of how this can be achieved in practice (Brier curves, ROC analysis).

In Section 7, we illustrate the observations of Sections 5 and 6 with a numerical example. Table 1 below demonstrates variance reduction through more powerful features both for the ML quantifier and a non-ML quantifier. Figure 1 below suggests that differences in efficiency between the ML quantifier and non-ML quantifiers may depend both on the information content (power) of the feature vector and on the true value of the positive class prevalence.

2. Related work

Prior probability shift is a special type of data set shift, see [11] for background information and a taxonomy of data set shift. In the literature, also other terms are used for prior probability shift, for instance 'global drift' [12] or 'label shift' [13].

The problem of estimating the test set prior class probabilities can also be interpreted as a problem to estimate the parameters of a 'mixture model' [14] where the component distributions are learnt on a training set. See [15] for an early work on the properties of the maximum likelihood (ML) estimator in this case. [16] revived the interest in the ML estimator for the unknown prior class probabilities in the test set by specifying the associated 'expectation maximisation' (EM) algorithm.

[17] proposed to take recourse to the notion of Fisher consistency as a criterion to identify completely unsuitable approaches to the quantification problem that do not have this property. [17] then proved Fisher consistency of the ML estimator under prior probability shift.

The ML approach has been criticised for its sometimes moderate performance and the effort and amount of training data needed to implement it. However, recently some researchers [9, 10] began to vindicate the ML approach. They focussed on the need to properly calibrate the posterior class probability estimate in order to improve the efficiency of the MLE for the positive class prevalence on the test set. Complementing the work of [9, 10], in this paper we study the role that constructing more powerful (or accurate) classifiers by selection of more appropriate features may play to reduce the variances of the related quantifiers.

In the following, we revisit the well-known asymptotic efficiency property of ML estimators in the special case of the MLE for prior class probabilities in the binary setting and investigate how it is impacted by the power (accuracy) of the classifier the MLE is based on.

3. Setting

We consider the binary prevalence quantification problem in the following setting:

- There is a training (or source) data set $(x_1, y_1), \ldots, (x_m, y_m) \in \mathfrak{X} \times \{0, 1\}$. It is assumed to be an i.i.d. sample of a random vector (X, Y) with values in $\mathfrak{X} \times \{0, 1\}$. The vector (X, Y) is defined on a probability space (Ω, P), the training (or source) domain. The elements ω of Ω are the instances (or objects). Each instance ω belongs to one of the classes 0 and 1, and its class label is $Y(\omega) \in \{0, 1\}$. In addition, each instance ω has features $X(\omega) \in \mathfrak{X}$. Often, $\mathfrak{X}$

is the d-dimensional Euclidian space such that accordingly X is a real-valued random vector. See Appendix B.1 of [18] for more detailed comments of how this setting avoids the logical problems that arise when feature vectors and instances are considered to be the same thing.

- Under the training distribution P, both the features $X(\omega)$ and the class labels $Y(\omega)$ of the instances are observed in a series of m independent experiments resulting in the sample $(x_1, y_1), \ldots, (x_m, y_m)$. The sample can be used to infer the joint distribution of X and Y under P, and hence, in particular, also the distribution of Y (the class distribution) under P.
- There is a test (or target) data set $z_1, \ldots, z_n \in \mathfrak{X}$. It is assumed to be an i.i.d. sample of the random vector X with values in $\mathfrak{X}$, under a probability measure Q on Ω that may be different to the training distribution P.
- Under the test distribution Q, only the features $X(\omega)$ of the instances are observed in a series of n independent experiments resulting in the sample $z_1, \ldots, z_n$. The sample can be used to infer the distribution of X under Q.
- The goal of quantification is to infer the distribution of Y under Q, based on the sample of features $z_1, \ldots, z_n$ generated under Q and on the joint sample of features and class labels $(x_1, y_1), \ldots, (x_m, y_m)$ generated under P. It is not possible to design a method for this inference without any assumption on the relation of P and Q.
- In this paper, we assume that P and Q are related by *prior probability shift*, in the sense that the class-conditional feature distributions are the same under P and Q, i.e. it holds that

$$P[X \in M \,|\, Y = i] \;=\; Q[X \in M \,|\, Y = i] \quad (1)$$

for $i \in \{0, 1\}$ and all measurable subsets M of $\mathfrak{X}$.

Denoting $P[Y = 1] = p$ and $Q[Y = 1] = q$, (1) implies that the distribution of the features X under P and Q respectively can be represented as

$$P[X \in M] = \quad (2a)$$
$$p\, P[X \in M \,|\, Y = 1] + (1 - p)\, P[X \in M \,|\, Y = 0],$$
$$Q[X \in M] = \quad (2b)$$
$$q\, P[X \in M \,|\, Y = 1] + (1 - q)\, P[X \in M \,|\, Y = 0],$$

for $M \subset \mathfrak{X}$. In the following, we assume that the components p, $P[X \in M \,|\, Y = 1]$ and $P[X \in M \,|\, Y = 0]$ can be perfectly estimated from the training sample $(x_1, y_1), \ldots, (x_m, y_m)$.

Basically, this means letting $m = \infty$ which obviously is infeasible. The assumption helps, however, to shed light on the importance of both maximum likelihood estimation and accurate classifiers for the efficient estimation of the unknown positive class prevalence q in the test data set.

4. The ML estimator for the positive class prevalence

Assume that the conditional distributions in (1) have positive densities f_i, $i = 0, 1$. Then the unconditional density of the features vector X under Q is

$$f^{(q)}(x) \;=\; (1 - q)\, f_0(x) + q\, f_1(x), \quad x \in \mathfrak{X}. \quad (3)$$

Hence the likelihood function

$$L_n(q) = L_n(q; z_1, \ldots, z_n)$$

for the sample $z_1, \ldots, z_n$ is given by

$$L_n(q) \;=\; \prod_{i=1}^{n} \big(q\,(f_1(z_i) - f_0(z_i)) + f_0(z_i)\big). \quad (4)$$

This implies for the first two derivatives of the log-likelihood with respect to q:

$$\frac{\partial \log L_n}{\partial\, q}(q) = \quad (5a)$$
$$\sum_{i=1}^{n} \frac{f_1(z_i) - f_0(z_i)}{q\,(f_1(z_i) - f_0(z_i)) + f_0(z_i)},$$
$$\frac{\partial^2 \log L_n}{\partial\, q^2}(q) = \quad (5b)$$
$$-\sum_{i=1}^{n} \left(\frac{f_1(z_i) - f_0(z_i)}{q\,(f_1(z_i) - f_0(z_i)) + f_0(z_i)} \right)^2 \leq 0.$$

We assume that there is at least one $j \in \{1, \ldots, n\}$ such that

$$f_1(z_j) \;\neq\; f_0(z_j). \quad (6)$$

Under (6), $q \mapsto \log L_n(q)$ is strictly concave in $[0, 1]$. Hence (see Example 4.3.1 of [19]) the equation

$$\frac{\partial \log L_n}{\partial\, q}(q) \;=\; 0 \quad (7)$$

has a solution $0 < q < 1$ if and only if

$$\frac{1}{n} \sum_{i=1}^{n} \frac{f_1(z_i)}{f_0(z_i)} > 1 \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{n} \frac{f_0(z_i)}{f_1(z_i)} > 1. \quad (8a)$$

This solution is then the unique point in $[0, 1]$ where $L_n(q)$ takes its absolute maximum value. By strict concavity of $\log L_n$, if (8a) is not true then either

$$\frac{1}{n} \sum_{i=1}^{n} \frac{f_1(z_i)}{f_0(z_i)} \leq 1 \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{n} \frac{f_0(z_i)}{f_1(z_i)} > 1, \quad (8b)$$

applies or

$$\frac{1}{n}\sum_{i=1}^{n}\frac{f_1(z_i)}{f_0(z_i)} > 1 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n}\frac{f_0(z_i)}{f_1(z_i)} \leq 1, \tag{8c}$$

holds. Under (8b), the unique maximum of $\log L_n$ in $[0, 1]$ lies at $q = 0$ while under (8c), the unique maximum of $\log L_n$ in $[0, 1]$ is taken at $q = 1$.

In summary, under the natural assumption (6), the likelihood function L_n of (4) has an absolute maximum in $[0, 1]$ at a unique point $q^* \in [0, 1]$. As a consequence, the maximum likelihood (ML) estimate $\hat{q}_n$ of the test set positive class prevalence q is well-defined by setting $\hat{q}_n = q^*$.

5. The Cramér-Rao bound for unbiased estimators

In the setting of Section 3, let $\widetilde{q}_n$ be any unbiased estimator of the positive class prevalence q under the test distribution Q, i.e. $\widetilde{q}_n = W_n(X_1, \ldots, X_n)$ for some function $W_n : \mathfrak{X}^n \to \mathbb{R}$ such that

$$E_Q[\widetilde{q}_n] = Q[Y = 1] = q,$$

where E_Q denotes the expected value under the mixture probability measure Q of (2b). Assume additionally that there are positive densities f_0 and f_1 of the class-conditional distributions of the feature vector X, such that the density $f^{(q)}$ of X under Q is given by (3).

If $(X_1, \ldots, X_n)$ is an i.i.d. sample from the distribution of the feature vector under the test distribution Q, then the variance of $\widetilde{q}_n$ is bounded from below by the inverse of the product of the Fisher information of the test distribution with respect to q and the size of the sample (Cramér-Rao bound, see, e.g., Corollary 7.3.10 of [20]):

$$\begin{aligned}\operatorname{var}[\widetilde{q}_n] &\geq \frac{1}{n\, E_Q\left[\left(\frac{\partial \log f^{(q)}}{\partial q}(X)\right)^2\right]} \\ &= \frac{1}{n\, E_Q\left[\left(\frac{f_1(X)-f_0(X)}{f^{(q)}(X)}\right)^2\right]}.\end{aligned} \tag{9}$$

Actually, since $\hat{q}_n$ from Section 4 is the ML estimator of q, (2b) presents not only a lower bound for the variances of unbiased estimators of q but also the large sample variance of $\hat{q}_n$ in the sense that $\sqrt{n}\,(\hat{q}_n - q)$ converges in distribution toward $\mathcal{N}\left(0, E_Q\left[\left(\frac{f_1(X)-f_0(X)}{f^{(q)}(X)}\right)^2\right]^{-1}\right)$, the normal distribution with mean 0 and variance $E_Q\left[\left(\frac{f_1(X)-f_0(X)}{f^{(q)}(X)}\right)^2\right]^{-1}$, the *asymptotic variance* of $\hat{q}_n$ (see, e.g., Theorem 10.1.12 of [20]). Note that the lower bound of (9) may not hold for $\hat{q}_n$ because it need not be unbiased.

In summary, if the densities of the class-conditional feature distributions are known, the ML estimator of q has asymptotically the smallest variance of all unbiased estimators of q on i.i.d. samples from the test distribution of the features. This is demonstrated in the two upper panels of Table 4 of [8] which shows on simulated data that confidence intervals for q – which are primarily driven by the standard deviations of the estimators – based on the ML estimator are the shortest if the training sample is infinite and the test sample is large.

6. The asymptotic variance of the ML estimator

Denote by $\eta_Q(x)$ the posterior positive class probability given $X = x$ under Q. Assume that the feature vector X under Q has a density that is given by (3). Then it holds that

$$\eta_Q(x) = \frac{q\, f_1(x)}{f^{(q)}(x)}. \tag{10}$$

This representation immediately implies the following result on the representation of the Fisher information mentioned above in the context of the Cramér-Rao bound in terms of the variance of η_Q.

Proposition 6.1. *If the feature vector X under Q has a density as specified by (3) then the Fisher information $E_Q\left[\left(\frac{\partial \log f^{(q)}}{\partial q}(X)\right)^2\right]$ of the distribution of X under Q with respect to q can be represented as follows:*

$$E_Q\left[\left(\frac{\partial \log f^{(q)}}{\partial q}(X)\right)^2\right] = \frac{\operatorname{var}_Q[\eta_Q(X)]}{q^2\,(1-q)^2}.$$

From Proposition 6.1 we obtain the following representation of the asymptotic variance of the ML estimator $\hat{q}_n$:

$$E_Q\left[\left(\frac{f_1(X)-f_0(X)}{f^{(q)}(X)}\right)^2\right]^{-1} = \frac{q^2\,(1-q)^2}{\operatorname{var}_Q[\eta_Q(X)]}. \tag{11}$$

Recall the following decomposition of the optimal Brier Score $BS_Q(X)$ for the problem to predict the class variable Y from the features X (under the test distribution Q):

$$\begin{aligned}BS_Q(X) &= E_Q\left[(Y - \eta_Q(x))^2\right] \\ &= \operatorname{var}_Q[Y] - \operatorname{var}_Q[\eta_Q(X)] \\ &= q\,(1-q) - \operatorname{var}_Q[\eta_Q(X)].\end{aligned} \tag{12}$$

In (12), the optimal Brier Score $BS_Q(X)$ is also called *refinement loss*, while $\mathrm{var}_Q[Y]$ and $\mathrm{var}_Q[\eta_Q(X)]$ are known as *uncertainty* and *resolution* respectively [21].

By (11), the asymptotic variance of the ML estimator $\hat{q}_n$ is reduced if a feature vector X with greater variance of $\eta_Q(X)$ is found (or if the Brier Score with respect to X decreases). Consider, for instance, a feature vector X' that is a function of the feature vector X, i.e. it holds that $X' = F(X)$ for some function $F : \mathfrak{X} \to \mathfrak{X}'$. Since F potentially maps different values $x \in \mathfrak{X}$ onto the same value $x' \in \mathfrak{X}'$ the amount of information carried by X' is reduced compared to the amount of information carried by X. Therefore, the approximation of the class label Y by regression on X' is less close than the approximation of Y by regression on X.

From this observation, it follows that $BS_Q(X') \geq BS_Q(X)$. This in turn implies by (12) and (11) for the asymptotic variances of the ML estimators $\hat{q}_n(X')$ and $\hat{q}_n(X)$ that

$$\mathrm{var}_Q[\hat{q}_n(X')] \;\leq\; \mathrm{var}_Q[\hat{q}_n(X)]. \tag{13}$$

Observe that $X' = F(X)$ also implies $BS_P(X') \geq BS_P(X)$, i.e. also under the training distribution P, the posterior positive class probability $\eta_P(X)$ based on X is a better predictor of Y than $\eta_P(X')$ which is based on X'. By the assumption underlying this paper, $BS_P(X')$ and $BS_P(X)$ are observable while $BS_Q(X')$ and $BS_Q(X)$ are not, because the class label Y is not observed in the test data set.

But does $BS_P(X') \geq BS_P(X)$ always imply $BS_Q(X') \geq BS_Q(X)$ and therefore also (13), thus generalising the implication "$X' = F(X) \Rightarrow$ (13)"?

This paper has no fully general answer to this question. Instead we can only point to alternative conditions on X' and X that imply both $BS_P(X') \geq BS_P(X)$ and $BS_Q(X') \geq BS_Q(X)$, but are weaker than $X' = F(X)$.

Brier curves: Recall the notion of *Brier curve* from [22] (with the slightly modified definition of [23]). If the Brier curve for $\eta_P(X')$ dominates the Brier curve for $\eta_P(X)$, then by item 6) of Proposition 5.2 and Proposition 4.1 of [23], it follows that $BS_P(X') \geq BS_P(X)$ and $BS_Q(X') \geq BS_Q(X)$ hold.

ROC analysis: Recall the notion of *Receiver Operating Characteristic (ROC)* as defined, for instance, in [24]. If the ROC for the density ratio associated with X dominates the ROC for the density ratio associated with X', then by Remark 5.4 and Proposition 4.1 of [23], it follows that $BS_P(X') \geq BS_P(X)$ and $BS_Q(X') \geq BS_Q(X)$ hold.

7. Example: Binormal model

In this section, we numerically compare the variance of the Sample Mean Matching (SMM) estimator $\hat{q}_{n,SMM}$ of class prevalences [2] and the Cramér-Rao bound of (9) (which is also the large sample variance of the ML estimator $\hat{q}_{n,ML}$ as specified in Section 4 above). In order to be able to do this, we take recourse to the univariate binormal model with equal variances of the class-conditional distributions: The two normal class-conditional distributions of the feature variable X are given by

$$X \,|\, Y = i \;\sim\; \mathcal{N}(\mu_i, \sigma^2), \qquad i = 0, 1, \tag{14a}$$

for conditional means $\mu_0 < \mu_1$ and some $\sigma > 0$. For the sake of simplicity, we choose

$$\mu_0 = 0, \quad \mu_1 > 0, \quad \sigma = 1. \tag{14b}$$

Greater values of μ_1 imply less overlap of the class-conditional feature distributions, corresponding to more powerful (or accurate) models. Or in other words, for greater values of μ_1, the feature variable X carries more information on the class label Y.

As stated in Section 3, we assume we are dealing with an infinitely large training sample and a test sample of size n. By Section 6, then for large n the variance of $\hat{q}_{n,ML}$ is approximately

$$\frac{1}{n}\, E_Q\left[\left(\frac{f_1(X) - f_0(X)}{f^{(q)}(X)}\right)^2\right]^{-1}, \tag{15}$$

where Q denotes the distribution underlying the test sample and f_0 and f_1 are the class-conditional feature densities – which are given for the purpose of this section by (14a). We evaluate the term given by (15) by means of one-dimensional numerical integration, making use of the R-function 'integrate' [25].

By Eq. (2) of [2], in the setting of this paper as specified in Section 3 above, the estimator $\hat{q}_{n,SMM}$ is given by the following explicit formula:

$$\begin{aligned} \hat{q}_{n,SMM} \;&=\; \frac{\frac{1}{n}\sum_{i=1}^n X_i - \mu_0}{\mu_1 - \mu_0} \\ &=\; \frac{1}{n\,\mu_1}\sum_{i=1}^n X_i. \end{aligned} \tag{16}$$

By (16), $\hat{q}_{n,SMM}$ is an unbiased estimator of the positive class prevalence q. For the variance of $\hat{q}_{n,SMM}$, we can refer to Theorem 3 of [7], case $n_{tr} = \infty$ in the notation of [7]. Observe that in the case of SMM and $n_{tr} = \infty$, it holds that the representation of the variance is exact, not only approximate. Hence we obtain

$$\begin{aligned} \mathrm{var}_Q[\hat{q}_{n,SMM}] \;&=\; \frac{\sigma^2 + q\,(1-q)\,(\mu_0^2 + \mu_1^2)}{n\,(\mu_1 - \mu_0)^2} \\ &=\; \frac{1}{n}\left(\frac{1}{\mu_1^2} + q\,(1-q)\right). \end{aligned} \tag{17}$$

Table 1
Illustration of the relations of model power, standard deviation of the SMM quantifier and large sample standard deviation of the ML quantifier. Sample size is 100, positive class prevalence in test set is 0.2. The parameter μ_1 is the expected sample mean conditional on the positive class, see (14a) and (14b).

μ_1	$AUC(\mu_1)$	σ_{SMM}	σ_{ML}
0.01	0.5028	10.0001	10.0000
0.05	0.5141	2.0004	2.0000
0.10	0.5282	1.0008	0.9999
0.25	0.5702	0.4020	0.3998
0.50	0.6382	0.2040	0.2000
1.00	0.7602	0.1077	0.1017
1.50	0.8556	0.0777	0.0710
2.00	0.9214	0.0640	0.0571
2.50	0.9615	0.0566	0.0498
3.00	0.9831	0.0521	0.0456
3.50	0.9933	0.0492	0.0432
4.00	0.9977	0.0472	0.0418
5.00	0.9998	0.0447	0.0405

In the model specified by (14a) and (14b), the classification power (or accuracy) is driven by the difference of the conditional means, i.e. by the mean conditional on the positive class μ_1. If we measure the power by the Area under the Curve (AUC, see for instance Section 6.1 of [24]) in order to obtain a measure which is independent of the class prevalences, AUC is a simple function of $\mu_1 = \mu_1 - \mu_0$:

$$AUC(\mu_1) \;=\; \Phi\left(\frac{\mu_1}{\sqrt{2}}\right), \tag{18}$$

with Φ denoting the standard normal distribution function.

From (18) und (17), it is clear that for fixed test sample size n the variance of $\hat{q}_{n,SMM}$ decreases when the power of the model increases. This is less obvious from (15) for the asymptotic variance of $\hat{q}_{n,ML}$ but if follows from (11) in that case.

Table 1 above illustrates these observations. In the table, we use the notation

$$\sigma_{SMM} = \sqrt{\mathrm{var}_Q[\hat{q}_{n,SMM}]}$$

and

$$\sigma_{ML} = \sqrt{\mathrm{var}_Q[\hat{q}_{n,ML}]}.$$

Table 1 suggests the following observations:

- At low levels of model power, small increases of power entail huge reductions of both the SMM variance and the ML large sample variance.
- The variance reductions become moderate or even low for moderate and high levels of model power (higher than 75% AUC).

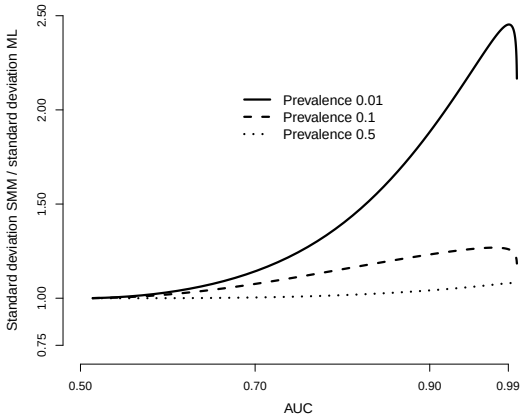

Figure 1: Ratios of standard deviations of SMM and ML estimates as function of AUC and positive class prevalence.

- The efficiency deficiency of the SMM estimator (as expressed by its variance) compared to the large sample ML variance varies and is much larger for higher levels of model power.

Figure 1 shows that the efficiency gain of the ML estimator compared to the SMM estimator – assuming an infinitely large training sample – does not only depend on the model power but also on the positive class prevalence. Indeed, Figure 1 suggests that a large efficiency gain is possible in the presence of a large difference in the prevalences of the two classes while the gain is rather moderate in the case of almost equal class prevalences.

8. Conclusions

In this paper, we have revisited the binary quantification problem, i.e. the problem of estimating a binary prior class distribution on the test data set when training and test distributions are different.

- Specifically, under the assumption of prior probability shift we have looked at the asymptotic variance of the maximum likelihood estimator (MLE).
- We have found that this asymptotic variance is closely related to the Brier score for the regression of the class label variable Y against the features vector X under the test set distribution. In particular, the asymptotic variance can be reduced by selection of a more powerful feature vector.
- At the end of Section 6, we have pointed out sufficient conditions and associated training criteria (Brier curves and ROC analysis) for minimising both the Brier score on the training data set and the Brier score on the test data set.

- These findings suggest methods to reduce the variance of the ML estimator of the prior class probabilities (or prevalences) on the test data set. Due to the statistical consistency of ML estimators, by reducing the variance of the estimator also its mean squared error is minimised.
- The large sample variance of the MLE associated with its asymptotic variance is identical to the Cramér-Rao lower bound for the variances of unbiased estimators of the prior positive class probability. Therefore, it seems likely that also other estimators benefit from improved performance of the underlying classifiers or feature vectors.

Indeed, results of a simulation study in [8] and theoretical findings in [7, 26] suggest that improving the accuracy of the base classifiers used for quantification helps to reduce not only the variances of ML estimators but also of other estimators. The example of the Sample Mean Match (SMM) estimator we have discussed in Section 7 supports this conclusion.

However, these findings must be qualified in so far as the observations made in this paper apply only to the case where both training data set and test data set are large. This is a severe restriction indeed as [5] pointed out the importance of quantification methods specifically in the case of small training data sets, for cost efficiency reasons.

Further research on developing efficient quantifiers for small or moderate training and test data set sizes therefore is highly desirable. A promising step in this direction has already been done by [27], with a proposal for the selection of the most suitable quantifiers for problems on data sets with widely varying sizes.

Acknowledgments

The author thanks four reviewers for their useful and supportive comments.

References

[1] P. González, A. Castaño, N. Chawla, J. D. Coz, A Review on Quantification Learning, ACM Comput. Surv. 50 (2017) 74:1–74:40.

[2] W. Hassan, A. Maletzke, G. Batista, Accurately Quantifying a Billion Instances per Second, in: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), 2020, pp. 1–10. doi:10.1109/DSAA49011.2020.00012.

[3] A. Moreo, A. Esuli, F. Sebastiani, QuaPy: A Python-Based Framework for Quantification, arXiv preprint arXiv:2106.11057, 2021.

[4] G. Forman, Counting Positives Accurately Despite Inaccurate Classification, in: European Conference on Machine Learning (ECML 2005), Springer, 2005, pp. 564–575.

[5] G. Forman, Quantifying counts and costs via classification, Data Mining and Knowledge Discovery 17 (2008) 164–206.

[6] A. Esuli, F. Sebastiani, A. Abbasi, Sentiment quantification, IEEE intelligent systems 25 (2010) 72–79.

[7] A. Vaz, R. Izbicki, R. Stern, Prior Shift Using the Ratio Estimator, in: A. Polpo, J. Stern, F. Louzada, R. Izbicki, H. Takada (Eds.), International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Springer, 2017, pp. 25–35.

[8] D. Tasche, Confidence intervals for class prevalences under prior probability shift, Machine Learning and Knowledge Extraction 1 (2019) 805–831.

[9] A. Alexandari, A. Kundaje, A. Shrikumar, Maximum Likelihood with Bias-Corrected Calibration is Hard-To-Beat at Label Shift Adaptation, in: International Conference on Machine Learning, PMLR, 2020, pp. 222–232.

[10] S. Garg, Y. Wu, S. Balakrishnan, Z. Lipton, A Unified View of Label Shift Estimation, in: H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 3290–3300.

[11] J. Moreno-Torres, T. Raeder, R. Alaiz-Rodriguez, N. Chawla, F. Herrera, A unifying view on dataset shift in classification, Pattern Recognition 45 (2012) 521–530.

[12] V. Hofer, G. Krempl, Drift mining in data: A framework for addressing drift in classification, Computational Statistics & Data Analysis 57 (2013) 377–391.

[13] Z. Lipton, Y.-X. Wang, A. Smola, Detecting and Correcting for Label Shift with Black Box Predictors, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 3122–3130.

[14] S. Frühwirth-Schnatter, Finite Mixture and Markov Switching Models: Modeling and Applications to Random Processes, Springer, 2006.

[15] C. Peters, W. Coberly, The numerical evaluation of the maximum-likelihood estimate of mixture proportions, Communications in Statistics – Theory and Methods 5 (1976) 1127–1135.

[16] M. Saerens, P. Latinne, C. Decaestecker, Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure, Neural Computation 14 (2001) 21–41.

[17] D. Tasche, Fisher Consistency for Prior Probability Shift, Journal of Machine Learning Research 18

(2017) 1–32. URL: http://jmlr.org/papers/v18/17-048.html.

[18] M.-J. Zhao, N. Edakunni, A. Pocock, G. Brown, Beyond Fano's Inequality: Bounds on the Optimal F-Score, BER, and Cost-Sensitive Risk and Their Implications, The Journal of Machine Learning Research 14 (2013) 1033–1090.

[19] D. Titterington, A. Smith, U. Makov, Statistical analysis of finite mixture distributions, Wiley New York, 1985.

[20] G. Casella, R. Berger, Statistical Inference, second ed., Duxbury Press, 2002.

[21] D. Hand, Construction and Assessment of Classification Rules, John Wiley & Sons, Chichester, 1997.

[22] J. Hernández-Orallo, P. Flach, C. Ferri, Brier Curves: A New Cost-Based Visualisation of Classifier Performance, in: Proceedings of the 28th International Conference on Machine Learning (ICML 2011), International Machine Learning Society, 2011, pp. 585–592.

[23] D. Tasche, Calibrating sufficiently, arXiv preprint arXiv:2105.07283, 2021.

[24] M. Reid, R. Williamson, Information, Divergence and Risk for Binary Experiments, Journal of Machine Learning Research 12 (2011) 731–817.

[25] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2019. URL: https://www.R-project.org/.

[26] A. Vaz, R. Izbicki, R. Stern, Quantification Under Prior Probability Shift: the Ratio Estimator and its Extensions, Journal of Machine Learning Research 20 (2019) 1–33. URL: http://jmlr.org/papers/v20/18-456.html.

[27] A. Maletzke, W. Hassan, D. dos Reis, G. Batista, The Importance of the Test Set Size in Quantification Assessment, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20), 2020, pp. 2640–2646.