Multiple-source cross-validation

Krzysztof J. Geras K.J.GERAS@SMS.ED.AC.UK
Charles Sutton CSUTTON@INF.ED.AC.UK
School of Informatics, University of Edinburgh

Abstract

Cross-validation is an essential tool in machine learning and statistics. The typical procedure, in which data points are randomly assigned to one of the test sets, makes an implicit assumption that the data are exchangeable. A common case in which this does not hold is when the data come from multiple *sources*, in the sense used in transfer learning. In this case it is common to arrange the cross-validation procedure in a way that takes the source structure into account. Although common in practice, this procedure does not appear to have been theoretically analysed. We present new estimators of the variance of the cross-validation, both in the multiple-source setting and in the standard iid setting. These new estimators allow for much more accurate confidence intervals and hypothesis tests to compare algorithms.

1. Introduction

Cross-validation is an essential tool in machine learning and statistics. The procedure estimates the expected error of a learning algorithm by running a training and testing procedure repeatedly on different partitions of the data. In the most common setting, data items are assigned to a test partition uniformly at random. This scheme is appropriate when the data are independent and identically distributed, but in modern applications this iid assumption often does not hold.

One common situation is when the data arise from multiple *sources*, each of which has a characteristic generating process. For example, in document classification, text that is produced by different authors, different organisations or of different genres will have different characteristics that affect classification (Blitzer et al., 2007; Craven et al., 1998). As another example, in biomedical data, such as EEG data (Mitchell et al., 2004), data items are associated with particular subjects, with large variation across subjects.

For data of this nature, a common procedure is to arrange the cross-validation procedure by source, rather than assigning the data points to test blocks randomly. The idea behind this procedure is to estimate the performance of the learning algorithm when it is faced with data that arise from a new source that has not occurred in the training data. We call this procedure *multiple-source cross-validation*. Although it is commonly used in applications, we are unaware of theoretical analysis of this cross-validation procedure.

This paper focuses on the estimate of the prediction error that arises from the multiple-source cross-validation procedure. We show that this procedure provides an unbiased estimate of the performance of the learning algorithm on a new source that was unseen in the training data, which is in contrast to the standard cross-validation procedure. We also analyse the variance of the estimate of the prediction error, inspired by the work of Bengio & Grandvalet (2003). Estimating the variance enables the construction of approximate confidence intervals on the prediction error, and hypothesis tests to compare learning algorithms.

We find that estimators of the variance based on the standard cross-validation setting (Grandvalet & Bengio, 2006) perform extremely poorly in the multiple-source setting, in some cases, even failing to be consistent if allowed infinite data. Instead, we propose a new family of estimators of the variance based on a simple characterisation of the space of possible biases that can be achieved by a class of reasonable estimators. This viewpoint yields a new estimator not only for the variance of the multiple-source cross-validation but for that of standard cross-validation as well.

On a real-world text data set that is commonly used for studies in domain adaptation (Blitzer et al., 2007), we

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

demonstrate that the new estimators are much more accurate than previous estimators.

2. Background

Cross-validation (CV) is a common means of assessing the performance of learning algorithms (Stone, 1974). Given a loss function $L(y, \hat{y})$ and a learning algorithm A that maps a data set D to a predictor $A^D : x \mapsto y$, CV can be interpreted as a procedure to estimate the expected prediction error $\mu = \text{EPE}(A) = \mathbb{E}[L(A^D(x), y)]$. The expectation is taken over training sets D and test instances (x, y).

In this paper we focus on k-fold CV. We introduce some notation to compactly describe the procedure. Denote the training set by $D = \{(x_i, y_i)\}_{i=1}^N$ and partition the data as $D = D_1 \cup \ldots \cup D_K$ with $|D_k| = M$ for all k. This means that $N = KM$. Let $\text{HO}(A, D_1, D_2)$ denote the average loss incurred when training on set D_1, and testing on set D_2.

In k-fold cross-validation, we first compute the average loss, $\text{HO}(A, D \backslash D_k, D_k)$, when training A on $D \backslash D_k$ and testing on D_k, for each $k \in \{1, \ldots, K\}$. We average these to get the final error estimate

$$\hat{\mu} = \text{CV}(A, D_{1:K}) = \frac{1}{K} \sum_{k=1}^{K} \text{HO}(A, D \backslash D_k, D_k). \quad (1)$$

If $(D_1, \ldots, D_K)$ is a random partition of D, we will refer to this procedure as *random cross-validation* (CVR) to differentiate it from the *multiple-source cross-validation* (CVS) that is the principal focus of this paper.

Notice that every instance in D is a test instance exactly once. For $(x_m, y_m) \in D_k$, denote this test error by the random variable $e_{km} = L(y_m, A^{D \backslash D_k}(x_m))$. Using this notation, $\hat{\mu} = \frac{1}{KM} \sum_k \sum_m e_{km}$, so the CV estimate is a mean of correlated random variables. For CVR, the e_{km}'s have a special exchangeability structure: For each k, the sequence $\mathbf{e}_k = (e_{k1}, \ldots, e_{kM})$ is exchangeable, and the sequence of vectors $\mathbf{e} = (\mathbf{e}_1, \ldots, \mathbf{e}_K)$ is also exchangeable.

Our goal will be to estimate the variance $\theta_{\text{CVR}} = \mathbb{V}[\hat{\mu}_{\text{CVR}}]$ of this and related CV estimators. If the examples in D are iid, Bengio & Grandvalet (2003) show that the variance θ_{CVR} can be decomposed into a sum of three terms. The exchangeability structure of the e_{km} implies that there are only three distinct entries in their covariance matrix: σ^2, ω and γ, where

$$\begin{aligned} \mathbb{V}[e_{ki}] &= \sigma^2, && \forall i \in \{1, \ldots, M\}, \forall k \in \{1, \ldots, K\}, \\ \text{Cov}[e_{ki}, e_{kj}] &= \omega, && \forall i, j \in \{1, \ldots, M\}, i \neq j, \forall k \in \{1, \ldots, K\}, \\ \text{Cov}[e_{ki}, e_{\ell j}] &= \gamma, && \forall i, j \in \{1, \ldots, M\}, \forall k, \ell \in \{1, \ldots, K\}, k \neq \ell. \end{aligned}$$

Applying the formula for the variance of the sum of correlated variables yields

$$\theta_{\text{CVR}} = \mathbb{V}[\hat{\mu}_{\text{CVR}}] = \frac{1}{KM}\sigma^2 + \frac{M-1}{KM}\omega + \frac{K-1}{K}\gamma. \quad (2)$$

This decomposition can be used to show that an unbiased estimator of θ_{CVR} does not exist. The reasoning behind this is the following. Because θ_{CVR} has only second order terms in e_{ki}'s, so would an unbiased estimator. Thus, if there existed such an estimator, it would be of the form $\hat{\theta} = \sum_{k,\ell} \sum_{ij} w_{k\ell} e_{ki} e_{\ell j}$. Coefficients $w_{k\ell}$ would be found by equating coefficients of σ^2, ω and γ in $\mathbb{E}\left[\hat{\theta}\right]$ and in θ_{CVR}. Unfortunately, the resulting system of equations has no solution, unless some assumptions about σ^2, ω or γ are made. In later work, Grandvalet & Bengio (2006) suggested several biased estimators of θ_{CVR} based on this variance decomposition and simplifying assumptions on σ^2, ω and γ. These are described in Table 2.

3. Multiple-source cross-validation

In many practical problems, the data arise from a number of different *sources* that have different generating processes. Examples of sources include different genres in document classification, or different patients in a biomedical domain. In these cases, often the primary interest is in the performance of a classifier on a new source, rather than over only the sources in the training data. This is essentially the same setting used in domain adaptation and transfer learning, except that we are interested in estimating the error of a prediction procedure rather than in developing the prediction procedure itself.

This situation can be modelled by a simple hierarchical generative process. We assume that each source $k \in \{1, \ldots, K\}$ has a set of parameters β_k that define its generative process and that the parameters $\beta_1, \ldots, \beta_K$ for each source are iid according to an unknown distribution. We assume that the source of each datum in the training set is known, i.e., each training example is of the form (y_m, x_m, k_m), where k_m is the index of the source of data item m. The data are then modelled as arising from a distribution $p(y_m, x_m | \beta_{k_m})$ that is different for each source.

The goal of cross-validation in this setting is to estimate the *out-of-source* error, i.e., the error on data that arises from a new source that does not occur in the training data. This error is

$$\text{OSE} = \mathbb{E}[L(A^D(x), y)], \quad (3)$$

where the expectation is taken with respect to training

sets D and testing examples (y, x) that are drawn from a new source, that is, $p(y, x) = \int p(y, x|\beta)p(\beta)d\beta$.

In the multiple-source setting, it is intuitively obvious that the standard cross-validation estimator of (1) is inappropriate. (We shall make this intuition precise in the next section.) Instead it is common practice to modify the CV procedure so that no source is split across test blocks, a procedure that we will call *multiple-source cross-validation*. Let S_k denote the set of all training points that were sampled from source k. Multiple-source cross-validation works exactly as standard CV, except that we partition the data as $D = S_1 \cup \ldots \cup S_K$ instead of assigning the training instances to blocks randomly. The resulting estimate of the error, which we denote by CVS, is

$$\hat{\mu}_{\text{CVS}} = \text{CVS}(A, S_{1:K}) = \frac{1}{K}\sum_{k=1}^{K} \text{HO}(A, D\backslash S_k, S_k).$$

The idea behind this procedure is that it accounts for the fact that we expect to be surprised by the new source in the test data, by using the information in the training data to simulate the effect of predicting on an unseen source. Although this estimator is commonly used in practice, we are unaware of previous work concerning its asymptotic or finite sample behaviour.

4. Analysis of multiple-source cross-validation

We give the mean (Section 4.1) and variance (Section 4.2) of the CVS estimator. The variance has an analogous decomposition to the CVR estimator, but the individual terms in the CVS variance behave differently from those in CVR, in a way that has a large impact on effective estimators of the variance. Finally, we present novel estimators of the variance of the CVS error estimate (Section 4.2).

4.1. Mean of the error estimate

First, we consider the expected value of both the CVS and the CVR error estimates, in the case where the data was in fact generated from the multiple-source process described in the previous section. The idea is to make clear what $\hat{\mu}_{\text{CVS}}$ is estimating, and to provide a more careful justification for the common use of $\hat{\mu}_{\text{CVS}}$ on multiple-source data.

Let $D = S_1 \cup \ldots \cup S_K$. Then it is easy to show using the exchangeability of $S_1, \ldots, S_K$ that the expected value of the estimate $\hat{\mu}_{\text{CVS}}$ is

$$\begin{aligned}\mathbb{E}\left[\hat{\mu}_{\text{CVS}}\right] &= \mathbb{E}_{(S_{1:K})}\left[\text{HO}(A, S_{1:(K-1)}, S_K)\right] \\ &= \mathbb{E}_{(S_{1:(K-1)},X,Y)}\left[L(A^{S_{1:(K-1)}}(X), Y)\right], \quad (4)\end{aligned}$$

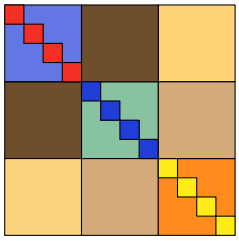

Figure 1. Diagram of the covariance matrix of the error variables for CVS. Blocks of the same colour represent blocks of variables with identical covariance.

where this notation indicates that the expectation is taken over $S_{1:(K-1)}$, X and Y. This is the expected error when the algorithm A is trained with data coming from $K-1$ sources and the test point is going to come from a newly sampled source. This is the out-of-source error (3), but on a slightly smaller training set than D.

On the other hand, consider the CVR estimate on the same data set D. Let $D_1 \cup \ldots \cup D_K = D$ be a random partition of D with $|D_k| = M$ for all k. The same symmetry argument which was used in (4) yields

$$\begin{aligned}\mathbb{E}\left[\hat{\mu}_{\text{CVR}}\right] &= \mathbb{E}_{(D_{1:K})}\left[\text{HO}(A, D_{1:(K-1)}, D_K)\right] \\ &= \mathbb{E}_{(D_{1:(K-1)},X,Y)}\left[L(A^{D_{1:(K-1)}}(X), Y)\right], \quad (5)\end{aligned}$$

which is formally similar to the above, but has the crucial difference that $D_{1:(K-1)}, X, Y$ have a different joint distribution. Here X and Y are drawn from the same family of K sources as the training data $D_{1:(K-1)}$. So $\mathbb{E}\left[\hat{\mu}_{\text{CVR}}\right]$ is the expected error of an algorithm trained with $M(K-1)$ data items from K sources, when the test point arises from one of the training sources.

Neither (4) or (5) is exactly the same as the out-of-sample error (3) that we want to estimate. Looking at the above, we see that $\hat{\mu}_{\text{CVS}}$ is biased for (3) because $\hat{\mu}_{\text{CVS}}$ uses slightly fewer training sources than OSE. On the other hand, $\hat{\mu}_{\text{CVR}}$ is biased for OSE because the training sets are slightly smaller and also because in $\hat{\mu}_{\text{CVR}}$ the training and test data are drawn from the same set of sources. However, if it is important to have a conservative estimate of the error, $\hat{\mu}_{\text{CVR}}$ has the advantage that its bias will tend to be negative, based on the expected effect of a smaller training set. (We verify this intuition experimentally in Section 6.)

4.2. Variance of the error estimate

Now we consider the variance of $\hat{\mu}_{\text{CVS}}$. The key difference between this case and the CVR case is that error variables e_{ki} and $e_{\ell j}$ are no longer identically distributed if $k \neq \ell$, as the corresponding data points

arise from different sources. There is still a block structure to the covariance matrix of $\mathbf{e}$, but it is more complex. Namely, the error variables have identical covariance structure within each source but not across sources. More formally,

$$\begin{aligned} \mathbb{V}\left[e_{ki}\right] &= \sigma_k^2, & \forall i \in \{1,\ldots,M\}, \\ \mathrm{Cov}\left[e_{ki}, e_{kj}\right] &= \omega_k, & \forall i,j \in \{1,\ldots,M\}, i \neq j, \\ \mathrm{Cov}\left[e_{ki}, e_{\ell j}\right] &= \gamma_{k\ell}, & \forall i,j \in \{1,\ldots,M\}, k \neq \ell. \end{aligned} \quad (6)$$

This covariance structure follows from the generating process described in Section 3, and is depicted graphically in Figure 1. The means of the error variables have an analogous structure, that is, $\mathbb{E}\left[e_{ki}\right] = \mathbb{E}\left[e_{kj}\right] := \mu_k$, which follows because the data points within each source are exchangeable.

This allows us to obtain a decomposition of the variance $\theta_{\mathrm{CVS}} = \mathbb{V}[\hat{\mu}_{\mathrm{CVS}}]$ of the CVS error estimate. This decomposition is

$$\theta_{\mathrm{CVS}} = \frac{1}{K^2 M}\sum_k \sigma_k^2 + \frac{M-1}{K^2 M}\sum_k \omega_k + \frac{1}{K^2}\sum_{k\neq l}\gamma_{kl}, \quad (7)$$

which again follows because θ_{CVS} is the variance of a sum of correlated random variables. Notice the difference to decomposition of θ_{CVR} in (2).

In the rest of this section we consider various estimates of θ_{CVS}. As θ_{CVS} depends quadratically on the variables e_{kl}, we will restrict our attention to estimators of the form $\hat{\theta} = \sum_{k,\ell}\sum_{i,j} w_{k\ell} e_{ki} e_{\ell j}$. That is, the estimator is a quadratic function of the variables $\mathbf{e}$, and as the error variables $\mathbf{e}_k$ within a single source k are exchangeable, we require such variables to be weighted equally. We refer to an estimator of this form as *quadratic*.

Every quadratic estimator can be written as a function of the empirical second moments of the error variables. If $\hat{\theta}$ is quadratic,

$$\hat{\theta} = \sum_k a_k s_k^{\sigma^2} + \sum_k b_k s_k^{\omega} + \sum_{k\neq l} c_{kl} s_{kl}^{\gamma},$$

where the $s_k^{\sigma^2}$, s_k^{ω} and s_{kl}^{γ}, are empirical moments

$$s_k^{\sigma^2} = \frac{1}{M}\sum_i e_{ki}^2, \quad s_k^{\omega} = \frac{1}{M(M-1)}\sum_{i\neq j} e_{ki} e_{kj},$$

$$s_{kl}^{\gamma} = \frac{1}{M^2}\sum_{i,j} e_{ki} e_{lj}.$$

4.2.1. No Unbiased Estimator

Using the decomposition (7), we can now follow the same reasoning as Bengio & Grandvalet (2003) to show that there is no unbiased estimator of θ_{CVS}. First, because $\theta_{\mathrm{CVS}} = \frac{1}{(KM)^2}\sum_{k,\ell}\sum_{i,j}\mathrm{Cov}\left[e_{ki}e_{\ell j}\right]$, an unbiased estimator must also be quadratic. The expectation of a quadratic estimator has the form

$$\mathbb{E}[\hat{\theta}] = \sum_k a_k(\sigma_k^2+\mu_k^2) + \sum_k b_k(\omega_k+\mu_k^2) + \sum_{k\neq l} c_{kl}(\gamma_{kl}+\mu_k\mu_l). \quad (8)$$

To get $\mathbb{E}[\hat{\theta}] = \theta_{\mathrm{CVS}}$, we need to match the coefficients in the equations (7) and (8), including the coefficients of the μ_k^2 and $\mu_k\mu_\ell$ terms that need to equal zero, since there clearly exist distributions such that $\mu_k\mu_l > 0$. This yields the system of equations

$$a_k = \frac{1}{K^2 M}, \;\; b_k = \frac{M-1}{K^2 M}, \;\; a_k + b_k = 0, \;\; c_{kl} = \frac{1}{K^2}, \;\; c_{kl} = 0. \quad (9)$$

Clearly, these equations are unsatisfiable, so no unbiased estimator of θ_{CVS} exists.

4.2.2. Naive Estimators

We can derive new estimators of the variance θ_{CVS} by following the reasoning used in the standard CV setting by Grandvalet & Bengio (2006). Unfortunately, as we will see, the resulting estimators perform poorly.

The idea is rather than attempting to define an estimator that is unbiased for all data distributions, instead define an estimator that is unbiased for a restricted class of data distributions, defined by assumptions on μ_k, σ_k^2, ω_k and γ_{kl}.

First, we restrict our attention to cases in which the mean prediction error across sources is the same, i.e., $\mu_k = \mu_l := \mu$. A quadratic estimator which is unbiased for this class of data distributions must satisfy the equations

$$\begin{aligned} a_k = \frac{1}{K^2 M}, \qquad b_k = \frac{M-1}{K^2 M}, \qquad c_{kl} = \frac{1}{K^2}, \\ \sum_k a_k + \sum_k b_k + \sum_{k\neq l} c_{kl} = 0. \end{aligned} \quad (10)$$

These equations also have no solution. However, if we further assume (following the reasoning of Grandvalet & Bengio (2006), even though it is unlikely to be a good assumption) that one of the sources $\tilde{k}$ has $\omega_{\tilde{k}} = 0$, then we no longer have to match the value of the coefficient $b_{\tilde{k}}$, removing one of the constraints. Solving the remaining equations yields

$$\begin{aligned} a_k = \frac{1}{K^2 M}, \qquad c_{kl} = \frac{1}{K^2}, \\ b_{\tilde{k}} = \frac{M-1}{K^2 M} - 1, \qquad b_k = \frac{M-1}{K^2 M}, \forall k \neq \tilde{k}. \end{aligned}$$

The corresponding estimator and its bias are shown in the first line of Table 1. Similarly, instead of assuming

$\omega_{\tilde{k}} = 0$, we could restrict ourselves to have a single $\gamma_{\tilde{k}\tilde{l}}$, or to have all of the ω_k's or γ_{kl}'s equal zero. This yields the remaining estimators in Table 1.

These estimators are appealingly simple, but, unfortunately, they have serious problems. The main problem is that their biases depend on the mean error μ_k of the sources. This is often an order of magnitude larger than the true variance θ_{CVS} that we are trying to predict. In particular, $\hat{\theta}^{\gamma}$, which is an analog of $\hat{\theta}_3$, the preferred estimator in the original CVR setting in the work of Grandvalet & Bengio (2006), is poor in this regard. Unlike the true variance, it does not even converge to 0 as $N \to \infty$, because the second term $\frac{1}{K^2}\left(\sum_{k=1}^{K}\mu_k^2 - \frac{\sum_{k\neq l}\mu_k\mu_l}{(K-1)}\right)$ in its bias does not converge to 0 in general.

The end result is that we can attempt to follow a similar strategy as in standard cross-validation to estimate the error variance, but doing so leads to extremely poor estimators. Instead, we will introduce new estimators that are specific to the multiple-source cross-validation setting.

4.2.3. Design of new estimators

Instead of the naive estimators from the previous section, we can derive better estimators by taking a different perspective. Instead of trying to design estimators that are unbiased for a restrictive set of scenarios, we consider the asymptotic behaviour of the bias decomposition of the quadratic variance estimator.

In the last section we saw that every quadratic estimator $\hat{\theta}$ has a bias of the form

$$\mathbb{E}[\hat{\theta} - \theta_{\text{CVS}}] = \sum_k \left(a_k - \frac{1}{K^2 M}\right)\sigma_k^2 + \sum_k \left(b_k - \frac{M-1}{K^2 M}\right)\omega_k + \sum_{k\neq l}\left(c_{kl} - \frac{1}{K^2}\right)\gamma_{kl} + \sum_k (a_k + b_k)\,\mu_k^2 + \sum_{k\neq l} c_{kl}\mu_k\mu_l.$$

Choosing the coefficients a_k, b_k, c_{kl} uniquely determines an estimator $\hat{\theta}$. Let us consider the relative magnitude of the terms in the bias decomposition. The error means μ_k^2 are usually a few of orders of magnitude larger than $\frac{1}{K^2 M}\sigma_k^2$, $\frac{M-1}{K^2 M}\omega_k$ and $\frac{1}{K^2}\gamma_{kl}$ as long as M is not too small. (We check this intuition experimentally in Section 6.) Therefore it seems sensible to require that all of the μ_k terms vanish. This implies that $a_k + b_k = 0$ and $c_{kl} = 0$ for all k, l.

Next, usually σ_k^2 is larger than ω_k or γ_{kl}, which suggests choosing $a_k = (K^2 M)^{-1}$ so that the σ_k^2 term vanishes as well. This yields the estimator

$$\hat{\theta}_A = \frac{1}{K^2 M}\left(\sum_k s_k^{\sigma^2} - \sum_k s_k^{\omega}\right). \tag{11}$$

The bias of this estimator is

$$\mathbb{E}[\hat{\theta}_A - \theta_{\text{CVS}}] = -\frac{1}{K^2}\sum_k \omega_k - \frac{1}{K^2}\sum_{k\neq l}\gamma_{kl}.$$

However, we can do better than this. Instead of requiring that the σ_k^2 term vanish, which is a very stringent requirement, we could instead require its coefficient to be $(KN)^{-1} = (K^2 M)^{-1}$, so that it becomes negligible for large N. This amounts to the requirment that $a_k = 2(K^2 M)^{-1}$, which results in the estimator

$$\hat{\theta}_B = \frac{2}{K^2 M}\left(\sum_k s_k^{\sigma^2} - \sum_k s_k^{\omega}\right). \tag{12}$$

This estimator is especially appealing because of the form of its bias, which is

$$\mathbb{E}[\hat{\theta}_B - \theta_{\text{CVS}}] = \frac{1}{K^2 M}\left(\sum_k \sigma_k^2 - (M+1)\sum_k \omega_k - M\sum_{k\neq l}\gamma_{kl}\right).$$

Now σ_k^2's and ω_k's are positive, and in practice γ_{kl}'s are almost always positive if M is not small. So what is appealing about this estimator is that the three terms have differing signs. In many situations, the difference between the first term and the second two will be of smaller magnitude than either of the three terms alone, causing $\hat{\theta}_B$ to be significantly less biased than the estimators in Table 1. We show this experimentally in Section 6.

5. New estimators of θ for CVR

It is actually possible to apply the same viewpoint from the previous section in order to provide new estimators for the variance of the standard cross-validation procedure. Previously known estimators of the variance θ_{CVR} (Table 2) were designed to be unbiased for a subclass of generating processes. By considering the bias decomposition directly, as in the previous section, we can design better estimators.

First, we give a proposition that describes the space of possible biases for quadratic estimators of θ_{CVR}.

Proposition 1. *Let $\hat{\theta} = \sum_{k,\ell}\sum_{ij} w_{k\ell} e_{ki} e_{\ell j}$ be a quadratic estimator. Then the bias $\mathbb{E}\left[\hat{\theta} - \theta_{CVR}\right]$ has the form $\alpha_1\sigma^2 + \alpha_2\omega + \alpha_3\gamma + \alpha_4\mu$. Also $\alpha_1 + \alpha_2 + \alpha_3 - \alpha_4 = -1$. Conversely, for every α_1, α_2, α_3, α_4 such that $\alpha_1 + \alpha_2 + \alpha_3 - \alpha_4 = -1$, there exists a quadratic estimator $\hat{\theta}$ with bias*

$$\mathbb{E}\left[\hat{\theta} - \theta_{CVR}\right] = \alpha_1\sigma^2 + \alpha_2\omega + \alpha_3\gamma + \alpha_4\mu^2.$$

Proof. For a quadratic estimator $\hat{\theta}$, let $a = M\sum_k w_{kk}$, let $b = M(M-1)\sum_k w_{kk}$ and $c = M^2\sum_k\sum_{l\neq k} w_{kl}$.

Table 1. Naive estimators of θ_{CVS} coming from solutions of simplified system of equations (10).

Unbiased if	**Estimator**	**Bias**
$\omega_{\tilde{k}} = 0$	$\hat{\theta}^{\omega}_{\tilde{k}} = \frac{1}{K^2 M}\sum_k s_k^{\sigma^2} + \frac{M-1}{K^2 M}\sum_k s_k^{\omega} + \frac{1}{K^2}\sum_{k\neq l} s_{kl}^{\gamma} - s_{\tilde{k}}^{\omega}$	$-\omega_{\tilde{k}} + \frac{1}{K^2}\sum_{k,l}\mu_k\mu_l - \mu_{\tilde{k}}^2$
$\gamma_{\tilde{k}\tilde{l}} = 0$	$\hat{\theta}^{\gamma}_{\tilde{k},\tilde{l}} = \frac{1}{K^2 M}\sum_k s_k^{\sigma^2} + \frac{M-1}{K^2 M}\sum_k s_k^{\omega} + \frac{1}{K^2}\sum_{k\neq l} s_{kl}^{\gamma} - \left(\frac{s^{\gamma}_{\tilde{k}\tilde{l}} + s^{\gamma}_{\tilde{l}\tilde{k}}}{2}\right)$	$-\gamma_{\tilde{k}\tilde{l}} + \frac{1}{K^2}\sum_{k,l}\mu_k\mu_l - \mu_{\tilde{l}}\mu_{\tilde{k}}$
$\forall_k \;\; \omega_k = 0$	$\hat{\theta}^{\omega} = \frac{1}{K^2 M}\sum_k s_k^{\sigma^2} + \frac{M-1-KM}{K^2 M}\sum_k s_k^{\omega} + \frac{1}{K^2}\sum_{k\neq l} s_{kl}^{\gamma}$	$-\frac{\sum_k \omega_k}{K} + \frac{1-K}{K^2}\left(\sum_k \mu_k^2 - \frac{\sum_{k\neq l}\mu_k\mu_l}{(K-1)}\right)$
$\forall_{k,l} \;\; \gamma_{kl} = 0$	$\hat{\theta}^{\gamma} = \frac{1}{K^2 M}\sum_k s_k^{\sigma^2} + \frac{M-1}{K^2 M}\sum_k s_k^{\omega} - \frac{1}{K^2(K-1)}\sum_{k\neq l} s_{kl}^{\gamma}$	$-\frac{\sum_{k\neq l}\gamma_{kl}}{K(K-1)} + \frac{1}{K^2}\left(\sum_k \mu_k^2 - \frac{\sum_{k\neq l}\mu_k\mu_l}{(K-1)}\right)$

Table 2. Estimators of θ for CVR suggested by Grandvalet & Bengio (2006).

Unbiased if	**Estimator**	**Bias**
$\mu = 0$	$\hat{\theta}_1 = \frac{1}{N}s_1 + \frac{M-1}{N}s_2 + \frac{N-M}{N}s_3$	μ^2
$\omega = 0$	$\hat{\theta}_2 = \frac{1}{N}s_1 - \frac{N+1-M}{N}s_2 + \frac{N-M}{N}s_3$	$-\omega$
$\gamma = 0$	$\hat{\theta}_3 = \frac{1}{N}s_1 + \frac{M-1}{N}s_2 - \frac{M}{N}s_3$	$-\gamma$
$\gamma = -\frac{M}{N-M}\omega$	$\hat{\theta}_4 = \frac{1}{N}s_1 - \frac{1}{N}s_2$	$-\frac{M}{N}\omega - \frac{N-M}{N}\gamma$

Taking expectations, we have

$$\mathbb{E}\left[\hat{\theta}\right] = a\sigma^2 + b\omega + c\gamma + (a+b+c)\mu^2.$$

Therefore the bias has the form

$$\mathbb{E}\left[\hat{\theta} - \theta_{\mathrm{CVR}}\right] = \left(a - \frac{1}{KM}\right)\sigma^2 + \left(b - \frac{M-1}{KM}\right)\omega + \left(c - \frac{K-1}{K}\right)\gamma + (a+b+c)\mu^2.$$

So the bias has the required form, with $\alpha_1+\alpha_2+\alpha_3-\alpha_4 = -1$. Conversely, let $\alpha_1+\alpha_2+\alpha_3-\alpha_4 = -1$. Then we can obtain an estimator $\hat{\theta}$ by setting $a = \alpha_1 + \frac{1}{KM}$, $b = \alpha_2 + \frac{M-1}{KM}$, $c = \alpha_3 + \frac{K-1}{K}$, and $d = a+b+c+1$. □

All of the existing estimators of θ_{CVR} (Table 2) have similar biases in the sense that the coefficients α_1, α_2, and α_3 are non-positive. But from the previous proposition, we know that there is a much larger set of coefficients available.

To design a new estimator, we observe that both in the results of Grandvalet & Bengio (2006) and our own results in Section 6, typically $\omega > \gamma$ and ω and γ are of similar magnitude. Therefore an estimator with bias of $\omega - 2\gamma$ will have smaller bias than the estimators from Table 2. Applying the previous result, this bias is achieved by the estimator

$$\hat{\theta}_5 = \frac{1}{N}s_1 + \frac{N+M-1}{N}s_2 - \frac{N+M}{N}s_3,$$

where s_1, s_2 and s_3 are the empirical moments

$$s_1 = \frac{1}{K}\sum_k s_k^{\sigma^2}, \quad s_2 = \frac{1}{K}\sum_k s_k^{\omega}, \quad s_3 = \frac{1}{K(K-1)}\sum_{k\neq l} s_{kl}^{\gamma}.$$

In the next section we show that its performance is superior in practice to the best previous estimator $\hat{\theta}_3$ given by Grandvalet & Bengio (2006).

6. Experiments

We evaluate the usefulness of $\hat{\mu}_{\mathrm{CVR}}$ and $\hat{\mu}_{\mathrm{CVS}}$ as estimators of out-of-source error and the estimators of θ_{CVR} and θ_{CVS} on a data set of product reviews from Amazon (Blitzer et al., 2007), which is frequently used as a benchmark data set in domain adaptation. The data contains reviews of products from 25 diverse domains corresponding to high-level categories on Amazon.com. The goal is to classify whether a review is positive or negative based on the review text. We take each product domain as being a separate source.

We experiment with the version of the data set which contains ten domains, each with 1000 positive and 1000 negative examples. We will use CV to estimate the prediction error of a simple naive Bayes classifier. (We have replicated these results with an SVM with a linear kernel.)

6.1. Bias and variance

First we measure the bias of $\hat{\mu}_{\mathrm{CVR}}$ and $\hat{\mu}_{\mathrm{CVS}}$ and compare them to the out-of-source error. To estimate this, we average over the set of training domains, the set of training instances, the test domain and the test instances. To get this, we first sample without replacement a given number of domains, keeping all of them but one as training domains and using the remaining one as a test domain. Given this, we sample 100 pairs consisting of training and test data sets, sampling data points from empirical distribution of the respective domains. Having these, we run CVR and CVS on the training domains, comparing the cross-validation estimate to the prediction on the out-of-source test set. Figure 2 shows this comparison as a function of the number of training sources K (left panel) and the

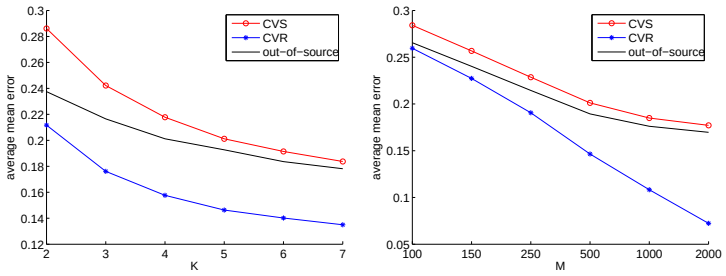

Figure 2. Values of $\hat{\mu}_{\text{CVR}}$ and $\hat{\mu}_{\text{CVS}}$ averaged over draws of training and test domains, compared to the true out-of-source error. Both plots were generated drawing domains 200 times.

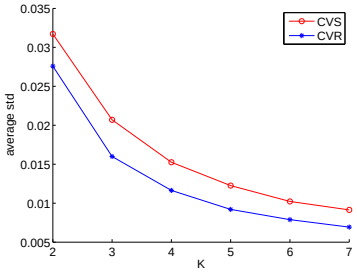

Figure 3. Standard deviation of $\hat{\mu}_{\text{CVR}}$ and $\hat{\mu}_{\text{CVS}}$ averaged draws of training and test domains. Experiment was done drawing domains 200 times.

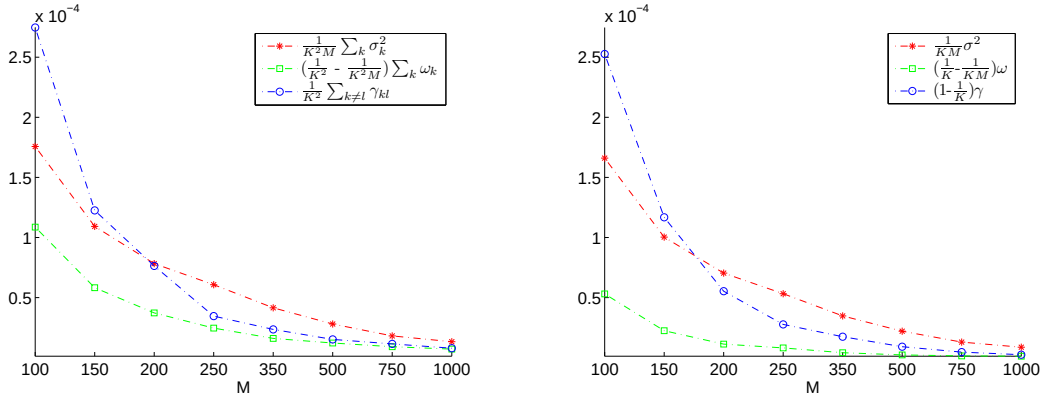

Figure 4. Decomposition of θ_{CVR} (right panel) and θ_{CVS} (left panel).

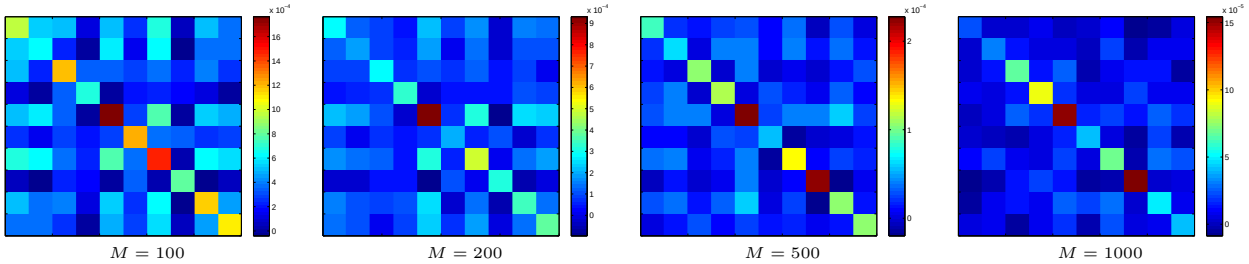

Figure 5. $\mathbb{V}[\mathbf{e}]$ for different M. σ_k^2's are typically a few orders of magnitude greater and were omitted here.

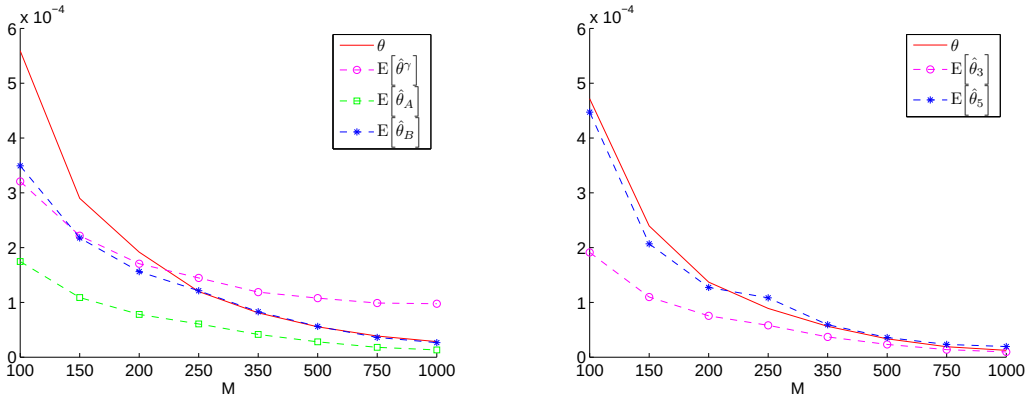

Figure 6. Comparison of estimators of θ_{CVR} (right panel) and θ_{CVS} (left panel).

number of points M per source (right panel). It can be clearly seen that CVS yields a better estimate of the out-of-source error than CVR. It is worth noting what happens when the number of training domains or the number of training data items gets larger. While CVS converges to the true out-of-source error, CVR converges to a different value. This confirms the analysis from Section 4.1. From these results, we can see that CVR yields an optimistic estimate of out-of-source error, even though the training set in each iteration of CV is smaller than the full training set. This is in contrast to CVS, which yields a more desirable pessimistic estimate.

Figure 3 shows the standard deviation of $\hat{\mu}_{\text{CVS}}$ and $\hat{\mu}_{\text{CVR}}$ averaged over the choice of training domains. To get this estimate, we have performed a similar procedure to Figure 2, i.e., for each draw of training domains, we sample 100 data sets, sampling data points from expirical distribution of the respective domains. Although the CVS estimate has higher variance, the variances are of the same order of magnitude and both tend to 0 for large data sets.

6.2. Variance decompositions and estimators of variance

In this section, we evaluate our new estimators of the variances θ_{CVR} and θ_{CVS}. To obtain these results, we used all domains as training domains. Estimating both θ_{CVR} and θ_{CVS} unbiasedly requires more than one independently sampled data set. To get them, we sample 1000 data sets, sampling from empirical distributions of each domain.

In the first experiment we estimated components of the decomposition of θ_{CVR} and θ_{CVS} (Figure 4). The quantities $\frac{1}{KM}\sigma^2$, $\frac{M-1}{KM}\omega$, $\frac{K-1}{K}\gamma$ and corresponding to them $\frac{1}{K^2M}\sum_k \sigma_k^2$, $\frac{M-1}{K^2M}\sum_k \omega_k$, $\frac{1}{K^2}\sum_{k\neq l}\gamma_{kl}$ have different magnitudes. Notice that $\frac{M-1}{KM}\omega$ is much larger than $\frac{M-1}{K^2M}\sum_k \omega_k$. It is not a surprise that CVS yields a strong positive correlation between errors made on points belonging to the same test block, which is not observed in CVR, where the test blocks are chosen randomly.

Secondly, we looked at $\mathbb{V}[\mathbf{e}]$ to see how much ω_k's and γ_{kl}'s vary (Figure 5). It can be seen that variation within γ_{kl}'s diminishes when M gets larger but variation within ω_k's does not change much.

Finally, we have tested estimators of θ_{CVR} and θ_{CVS}, which we have suggested in the earlier sections. The results are in Figure 6. For CVS, we compare $\hat{\theta}^{\gamma}$, $\hat{\theta}_A$ and $\hat{\theta}_B$. As expected, $\hat{\theta}^{\gamma}$ does not converge to 0 and grossly overestimates θ_{CVS} for large values of M. The new estimator $\hat{\theta}_A$ is closer to θ_{CVS} than $\hat{\theta}^{\gamma}$ for large values of M but its bias is consistently optimistic. On the other hand, for small M, $\hat{\theta}_B$ is not more optimistic than $\hat{\theta}^{\gamma}$ and for large M, while $\hat{\theta}^{\gamma}$ becomes very pessimistic, $\hat{\theta}_B$ has a negligible bias. Similarly, for the standard CVR setting, our new estimator $\hat{\theta}_5$ has lower bias than the previous estimators suggested by Grandvalet & Bengio (2006), being almost unbiased even for small M.

7. Related work

Various different versions of cross-validation have been analysed previously (Bengio & Grandvalet, 2003; Arlot & Celisse, 2010; Nadeau & Bengio, 2003; Markatou et al., 2005), but to our knowledge multiple-source cross-validation has not been previously analysed.

The idea of learning from a number of sources dates back to Caruana (1997) and Thrun (1996). A related issue in the context of covariate shift was suggested by Sugiyama et al. (2007), however, the resulting importance weights are difficult to estimate in practice (but see Gretton et al. (2009) for some work in this direction).

Rakotomalala et al. (2006) investigate the multiple-source cross-validation procedure empirically, but they do not perform theoretical analysis or present any estimators of θ_{CVS}. Work in domain adaptation also has considered the problem of bounding the prediction error when the training and test distribution have a different source (Ben-David et al., 2010). Unfortunately, the resulting bounds are too loose to be used for confidence intervals.

8. Conclusions

We have considered a cross-validation procedure for the multiple-source setting. We show that the bias of this procedure is better suited to this setting. We have presented several new estimates of the variance of the error estimate, both for the multiple-source cross-validation procedure and for the standard cross-validations setting, which perform well empirically.

Acknowledgments

The authors would like to thank Amos Storkey, Vittorio Ferrari, Simon Rogers and the anonymous reviewers for insightful comments on this work. This work was supported by the Scottish Informatics and Computer Science Alliance (SICSA).

References

Arlot, S. and Celisse, A. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. Wortman. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

Bengio, Y. and Grandvalet, Y. No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research*, 5:1089–1105, 2003.

Blitzer, J., Dredze, M., and Pereira, F. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *In Proceedings of Association of Computational Linguistics*, 2007.

Caruana, R. Multi-task learning. *Machine Learning*, 28:41–75, 1997.

Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., and Slatteryy, S. Learning to extract symbolic knowledge from the world wide web. In *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI-98)*, 1998.

Grandvalet, Y. and Bengio, Y. Hypothesis testing for cross-validation. Technical Report 1285, Département d'informatique et recherche opérationnelle, Université de Montréal, 2006.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. &. Covariate shift by kernel mean matching. In Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N.D. (eds.), *Dataset shift in machine learning*, pp. 131–160. The MIT Press, 2009.

Markatou, M., Tian, H., Biswas, S., and Hripcsak, G. Analysis of variance of cross-validation estimators of the generalization error. *Journal of Machine Learning Research*, 6:1127–1168, 2005.

Mitchell, T., Hutchinson, R., , Niculescu, R., Pereira, F., Wang, X., Justl, M., and Newman, S. Learning to decode cognitive states from brain images. *Machine Learning*, 57(1):145–175, 2004.

Nadeau, C. and Bengio, Y. Inference for the generalization error. *Machine Learning*, 52:239–281, 2003.

Rakotomalala, R., Chauchat, J.-H., and Pellegrino, F. Accuracy estimation with clustered dataset. In *In Proceedings of Australasian conference on Data mining and analytics*, 2006.

Stone, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147, 1974.

Sugiyama, M., Krauledat, M., and Müller, K.-R. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

Thrun, S. Is learning the n-th thing any easier than learning the first? In *In Advances in Neural Information Processing Systems*, 1996.