The Canadian Journal of Statistics
Vol. xx, No. yy, 20??, Pages 1–25
La revue canadienne de statistique

Partial order relations for classification comparisons

Lo-Bin Chang[1]*

[1]*The Ohio State University*

Key words and phrases: Partial Order; Classification Comparison; Cross Study Validation; Labeling error ; Outlier ; Bayes error rate; Receiver Operating Characteristic.

MSC 2010: Primary 62C05; secondary 62C07

Abstract: The Bayes classification rule offers the optimal classifier, minimizing the classification error rate, whereas the Neyman-Pearson lemma offers the optimal family of classifiers to maximize the detection rate for any given false alarm rate. These motivate studies on comparing classifiers based on similarities between the classifiers and the optimal. In this paper, we define partial order relations on classifiers and families of classifiers, based on rankings of rate function values and rankings of test function values, respectively. Each partial order relation provides a sufficient condition, which yields better classification error rates or better performance on the receiver operating characteristic (ROC) analysis. Various examples and applications of the partial order theorems are discussed to provide comparisons of classifiers and families of classifiers, including the comparison of cross-validation methods, training data that contains outliers, and labeling errors in training data.

Résumé: Insérer votre résumé ici. We will supply a French abstract for those authors who can't prepare it themselves.

1. INTRODUCTION

Over the last two decades as methodologies for building effective classifiers has continued to grow, interest in comparing the accuracy of these classification paradigms has become increasingly important across many fields of research (e.g., Lim, Loh, & Shih, 2000; Wu et al., 2003; Li et al., 2014; Dudoit, Fridlyand, & Speed, 2002). The Bayes classification rule and the Neyman-Pearson lemma provide the optimal classifier and the optimal family of classifiers respectively, and allow for theoretical studies on classification comparison (e.g., Lee & Wang, 2015; Chang et al., 2011). Nonetheless, most theoretical studies examining the comparison of classification methods are often overly restrictive and largely unexplored. This is likely due to the complicated analyses required to compute the classification error.

Motivated by the comparison of the *randomized cross validation* (RCV) and the *cross-study validation* (CSV) in Chang & Geman (2015), we define a partial order relation for comparison of classification error rates, which allows us to generalize the first theorem in the paper from 2-class classification to multi-class classification. Note that, without using the partial order relation, extending the original proof to multi-class classification problems is unfeasible.

Considering classifiers of the form $H(x) = \arg\max_{i\in\{1,2,\cdots,m\}} h_i(x)$, this paper proposes a partial order relation in which the order is defined by comparing, for every given x, the rankings of rate function values $\{h_i(x) : i = 1, 2, \cdots, m\}$ of classifiers with that of the optimal Bayes

* *Author to whom correspondence may be addressed.*
E-mail: lobinchang@stat.osu.edu

classifier. According to the partial order relation, a classifier $H_1 \leq H_2$ (another classifier) if H_1 ranks any pair of rate functions (i.e. $h_i(x), h_j(x)$ for any i, j and any x) more consistently with the oracle (Bayes classifier) than H_2 does. As demonstrated in Theorem 1, the partial order is consistent with the order of classification error rates, providing a sufficient condition for perturbing classifiers toward the Bayes optimal classifier (i.e. the *minmum* of the relation).

For comparing families of classifiers of the forms

$$H_t(x) = \begin{cases} 1 & \text{if } r(x) \geq t, \\ 2 & \text{if } r(x) < t, \end{cases}$$

a different partial order relation is proposed, in which the order is defined by comparing the rankings of test function values $\{r(x) : \text{for all } x\}$ of the families with that of the optimal families of classifiers (i.e. Neyman-Pearson classifiers). According to this partial order relation, a family is smaller (better) than another if the former ranks any pair of test function values (i.e. $r(x_1), r(x_2)$ for any x_1, x_2) more consistently with the oracle (i.e. the family of Neyman-Pearson classifiers) than the latter does. Similarly, this partial order relation provides a sufficient condition for perturbing classifiers toward the family of optimal Neyman-Pearson classifiers (the *minmum* of the relation). As demonstrated in Theorem 3, this partial order relation is consistent with comparing the performances of the receiver operating characteristic (ROC) curves.

In the next section, we discuss the definition of the partial order relation on classifiers and provide theoretical results and various examples, including the comparison of cross-validation methods, training sets that contain outliers, and the effects of labeling errors. Discussion of a partial order relation on families of classifiers and the ROC analysis, including applications for labeling errors and outlier contaminated models is included in Section 3. Numerical experiments are provided to demonstrate the properties of the RCV and CSV in Section 4. Discussion and open questions are included in Section 5. Proofs of theorems are in Appendix.

2. A PARTIAL ORDER RELATION ON CLASSIFIERS

Consider an m-class classification problem. Assume that the observed data are continuous and drawn from classes with possibly different d-dimensional densities (everything works similarly for discrete distributions). Let Y be a random integer in $\{1, \cdots, m\}$ representing the class and $p(i) = P(Y = i)$ be the prior probability of class i. Let X be a random vector in $\mathbf{R}^d$ representing the observation and $f(x|i)$ be the probability density of X given class $Y = i$.

For any classifier $H(x)$, let $h_i(x)$ be the $rate$ function for class i such that

$$H(x) = \arg \max_{i \in \{1,2,\cdots,m\}} h_i(x)$$

if the $argmax$ returns a single index, and that $H(x)$ otherwise is an integer from the set $\arg\max_{i \in \{1,2,\cdots,m\}} h_i(x)$. Let $\mathrm{Err}(H)$ be the classification error rate associated with $H(x)$:

$$\mathrm{Err}(H) = P(H(X) \neq Y) = \sum_{i=1}^{m} p(i) P_i(H(X) \neq i)$$

where $P_i(A) = \int_A f(x|i)dx$ for any measurable set A. The optimal classifier $H^o(x)$, which yields the Bayes error rate is defined as follows:

$$H^o(x) = \arg \max_{i \in \{1,2,\cdots,m\}} h_i^o(x),$$

where $h_i^o(x) = p(i)f(x|i)$. Notice that if $\arg\max_{i\in\{1,2,\cdots,m\}} h_i^o(x)$ returns a set of indices we can arbitrarily assign the class $H^o(x)$ from the set without affecting the error rate. Then we can define a partial order relation as follows to characterize the relative similarity to the optimal classifier.

Definition. *Two classifiers $\bar{H}(x)$ and $H(x)$ are of the relation $\bar{H} \leq H$, if there exist two sets of rate functions $\{\bar{h}_i(x) : i = 1, \cdots, m\}$ and $\{h_i(x) : i = 1, \cdots, m\}$ with which*

$$\bar{H}(x) = \arg\max_{i\in\{1,2,\cdots,m\}} \bar{h}_i(x), \quad H(x) = \arg\max_{i\in\{1,2,\cdots,m\}} h_i(x)$$

such that for any distinct $i, j \in \{1, 2, \cdots, m\}$ and $x \in \mathbf{R}^d$ we have $\bar{h}_i(x) > \bar{h}_j(x)$ whenever $h_i(x) \geq h_j(x)$ and $h_i^o(x) > h_j^o(x)$.

Consequently, the relation defined above is a partial order relation for which $H^o(x)$ is a minimum. Since $p(i)f(x|i) \propto f(i|x)$, the rate function h_i^o can be defined by $h_i^o(x) = f(i|x)$. Thus, this partial order relation can still be used in regression analysis or discriminative analysis when only the conditional distribution (posterior distribution) $f(i|x)$ is available or the corresponding rate function is proportional to $f(i|x)$. The following theorem shows that this relation is also a partial order relation associated with the classification error rate. The proof is in the appendix.

Theorem 1. *If $\bar{H} \leq H$, then the classification error rate associated with $\bar{H}(x)$ is less than or equal to that associated with $H(x)$ [i.e. $\mathrm{Err}(\bar{H}) \leq \mathrm{Err}(H)$].*

By this theorem, the partial order relation $\bar{H} \leq H$ provides a sufficient condition that ensures $\bar{H}$ has a better classification performance. Note that verifying this condition requires the underlying distributions or rate functions of the optimal classifier, which are usually unavailable in practice. However, many characteristics and theoretical results can be discovered by applying this partial order theorem, even when the underlying distributions and the optimal rate functions are unknown. In particular, by assuming perfect training without estimation errors, we can disregard the issues related to estimation and focus on the methodologies and other characteristics. The following subsections explore several examples and applications under various training scenarios to demonstrate the utility of the theorem.

2.1. Applications on Cross-Study Validation

"Study effects" are recently studied by comparing (ordinary) randomized cross-validation (RCV) with cross-study validation (CSV). For example, Ma et al. (2014) conducted a series of experiments to compare the variations of these two validation methods to demonstrate the impact of study effects. Chang & Geman (2015) proposed a statistical formulation, under which the theoretical results (three theorems) for two-class classification problems were established. In this subsection, we use the partial order theorem to extend the first theorem in Chang & Geman (2015) to multi-class classification problems.

Let the observed data be assembled from n equal-sized sources or "studies" (say $z_1, \cdots, z_n$), each consisting of samples from each of m classes. The populations among the studies could be heterogeneous so that the samples across the studies far from identically distributed. The RCV is a standard approach to estimating the error rate in which we train the classifier on the pooled data excluding a random subset (e.g. ten percent), test on the left-out subset, repeat the procedure and average the results. Nevertheless, in CSV, we leave each study out in turn, train on the other $n - 1$ studies, test on the left-out study and average the results. Of these two validation methods, which yields a larger error rate? Next, we show that the analytic error rate of CSV is larger or at least equal to that of RCV, in support of our intuition.

Now assume that the samples from each study and each class are continuously distributed. Let $p_z(k)$ be the prior class probabilities given study z for $k = 1, \cdots, m$, and $f_z(x|k)$ be class-conditional densities of X given class k and study z. Focusing on the influence of study effects to validation methods, we assume that the exact densities can be estimated in training procedures. Therefore, in the RCV training, we get the following mixture of n densities as the learned density for k-th class:

$$f_{z_{1:n}}(x|k) \equiv \frac{1}{\sum_{i=1}^{n} p_{z_i}(k)} \sum_{i=1}^{n} p_{z_i}(k) f_{z_i}(x|k), \tag{1}$$

where $k = 1, \cdots, m$. As per Chang & Geman (2015), the analytic error rate associated with RCV is the Bayes error rate associated with the m mixture densities:

$$\begin{aligned} & e_{RCV}(z_1, ..., z_n) \\ &= \sum_{k=1}^{m} \frac{\sum_{i=1}^{n} p_{z_i}(k)}{n} \times P\left(\arg \max_{t \in \{1,2,\cdots,m\}} \sum_{i=1}^{n} p_{z_i}(t) f_{z_i}(X|t) \neq k \middle| X \sim f_{z_{1:n}}(x|k) \right). \end{aligned}$$

Similarly, in the CSV training, we obtain the following mixture of $n-1$ densities as the learned density for the k-th class when z_j is the left-out study:

$$\frac{1}{\sum_{i \neq j} p_{z_i}(k)} \sum_{i \neq j} p_{z_i}(k) f_{z_i}(x|k), \tag{2}$$

where $k = 1, \cdots, m$. Thus, the analytic error rate associated with CSV is the average of n cross-study error rates:

$$\begin{aligned} & e_{CSV}(z_1, ..., z_n) \\ &= \frac{1}{n} \sum_{j=1}^{n} \left\{ \sum_{k=1}^{m} p_{z_j}(k) \cdot P\left(\arg \max_{t \in \{1,2,\cdots,m\}} \sum_{i \neq j} p_{z_i}(t) f_{z_i}(X|t) \neq k \middle| X \sim f_{z_j}(\cdot|k) \right) \right\}. \end{aligned}$$

We also have the following inequality for m-class classification, which is consistent with our intuition and the experimental results in Ma et al. (2014) where the proof is in the appendix.

Theorem 2. *For $n \geq 2$, $e_{CSV}(z_1, ..., z_n) \geq e_{RCV}(z_1, ..., z_n)$.*

Remark: The proof of the first theorem in Chang & Geman (2015) utilizes a key lemma, which can be proven effortlessly using the partial order theorem, even for the generalization to multi-class classification problems. However, extending the original proof of the lemma to multi-class classification problems is unfeasible.

2.2. Partial order relations for different training scenarios

2.2.1. A training set that contains outliers

We consider a scenario that the conditional densities given classes are estimated based on a training set that contains outliers generated from different distributions (say $\hat{f}(x|i)$'s). Therefore, the underlying densities are equal to $\bar{f}(x|i) = \epsilon \hat{f}(x|i) + (1-\epsilon) f(x|i)$, $\forall i = 1, \cdots, m$, where $\epsilon \in [0, 1]$ is the weight of outlier population. Denote the Bayes classifier associated with the $\bar{f}(x|i)$ functions as

$$H_{\bar{f}}(x) = \arg \max_{i \in \{1,2,\cdots,m\}} p(i) \bar{f}(x|i).$$

If $p(i)\hat{f}(x|i) \geq p(j)\hat{f}(x|j)$ and $p(i)f(x|i) > p(j)f(x|j)$, then $p(i)\bar{f}(x|i) > p(j)\bar{f}(x|j)$. Thus, we have $H_{\bar{f}} \leq H_{\hat{f}}$ by the partial order theorem. Similarly, we can show that $\mathrm{Err}(H_{\bar{f}})$ is nondecreasing in $\epsilon \in [0,1]$ and converges to the optimal error rate $\mathrm{Err}(H^o)$ as ϵ tends to 0.

The perturbed densities $\bar{f}(x|i) = \epsilon\hat{f}(x|i) + (1-\epsilon)f(x|i)$ can also be viewed as ϵ-contamination models in robustness analysis (e.g. He, Simpson, & Portnoy, 1990; Zio & Guarnera, 2013; Wellmann & Gather, 1999; Huber, 1981; Maronna, Martin, & Yohai, 2006). This property demonstrates that the classification error rate is nondecreasing in the contamination level ϵ. Hence, the monotonicity property agrees with robustness research and data mining in that incorporating less contaminated data or more current (more accurate) data in training yields better classifiers.

2.2.2. Labeling errors in a training set

The traditional training sets are assumed to be of correct labels (classes). However, accounting for inevitable mislabeling from human labeling mistakes, lack of information, or communication noise, the impact of label noise has attracted much attention for many years. Many articles focus on modifying learning methods to eliminate the influence of the label noise (e.g. Brodley & Friedl, 1999; Lawrence & Scholkopf, 2001; Leung, Song, & Zhang, 2011; Bootkrajang & Kaban, (2012); Scott, Blanchard, & Handy, 2013; Natarajan et al., 2013; Frenay & Verleysen, 2014). In this subsection, we study the performance of the Bayes classifiers learned based on training sets with labeling errors.

Consider an m-class classification problem. Assume that each label, say $Y = i$, in a training set has been independently flipped with probability $(m-1)\alpha_i$ to any of the other $(m-1)$ classes, each with probability α_i. Let $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_m)$ and assume $\alpha_i \leq 1/(m-1)$ for all i. Then, as the size of the training set tends to infinity, we assume that asymptotically the following distributions are learned: *Prior distribution,* $\tilde{p}_\alpha(i) = [1-(m-1)\alpha_i]p(i) + \sum_{s\neq i}\alpha_s p(s)$ and *Conditional data distribution,*

$$\tilde{f}_\alpha(x|i) = \frac{1}{\tilde{p}_\alpha(i)}\Big\{[1-(m-1)\alpha_i]p(i)f(x|i) + \sum_{s\neq i}\alpha_s p(s)f(x|s)\Big\}. \tag{3}$$

The Bayes classifier associated with the above distributions is as follows:

$$\begin{aligned}\tilde{H}_\alpha(x) &= \arg\max_{i\in\{1,2,\cdots,m\}} \tilde{p}_\alpha(i)\tilde{f}_\alpha(x|i) \\ &= \arg\max_{i\in\{1,2,\cdots,m\}} [1-(m-1)\alpha_i]p(i)f(x|i) + \sum_{s\neq i}\alpha_s p(s)f(x|s). \end{aligned} \tag{4}$$

Therefore the classification error rate of $\tilde{H}_\alpha(x)$ is a function of $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_m)$. Using the partial order theorem, we discovered the following partial order relation in terms of the maximum labeling error rate, denoted by $\alpha_{(1)} = \max_{1\leq i\leq n}\alpha_i$, where the proof is straightforward and therefore skipped.

Proposition 1. *Fix non-maximum α_i values and regard the classification error rate as a function of $\alpha_{(1)}$. Then, $\mathrm{Err}(\tilde{H}_\alpha)$ is nondecreasing in $\alpha_{(1)} \in [\alpha_{(2)}, 1/m)$, where $\alpha_{(2)}$ is the second largest α_i.*

Notice that we can show that when $\alpha_1 = \alpha_2 = \cdots = \alpha_m < \frac{1}{m}$ (i.e., uniform labeling error), the classification error rate is equal to the optimal classification error rate. Hence, reducing the non-maximum labeling error rates may result in worse classification because it might increase the non-uniformity of label noise. For example, for a fixed $\delta \in (0, 1/m)$, let $\alpha_1 = \delta - \epsilon$, and

$\alpha_i = \delta$ for all $i = 2, \cdots, m$, and assume $\mathrm{Err}(\tilde{H}_\alpha) \equiv q(\epsilon)$ is a continuous function of ϵ. Then it converges to the optimal classification error rate $\mathrm{Err}(H^o)$ when ϵ tends to zero. If $q(\epsilon_0) > \mathrm{Err}(H^o)$ for some $\epsilon_0 \in (0, \delta)$, then $q(\epsilon)$ must be increasing somewhere in $(0, \epsilon_0)$. Thus, reducing the first labeling error rate α_1 (i.e., increasing ϵ) could increase the classification error rate.

In addition, the proposition can be generalized to regression models by replacing $p(i)f(x|i)$ and $p(s)f(x|s)$ with $f(i|x)$ and $f(s|x)$ respectively in equation (4). Furthermore, other partial order relations in terms of labeling errors may also be discovered using the partial order theorem.

3. A PARTIAL ORDER RELATION ON FAMILIES OF CLASSIFIERS

Consider a two-class classification problem where class 1 is abnormal and class 2 is normal (e.g. cancer versus no cancer). Another common approach to comparing classification performance or powers of statistical hypothesis testing is to compare the ROC performances ($i.e.$ given a fixed level of false alarm rate we compare the detection rates or discovery rates). Consider the family of classifiers $\{H_t(x)\}$ of the following form associated with a test function $r(x)$:

$$H_t(x) = \begin{cases} 1 & \text{if } r(x) \geq t, \\ 2 & \text{if } r(x) < t. \end{cases}$$

In particular, the optimal family of classifiers, denoted by $\{H_t^o(x)\}$ uses the likelihood ratio function $r^o(x) = \frac{f(x|1)}{f(x|2)}$ as the test function where $f(x|1), f(x|2)$ are conditional densities for class 1 and class 2, respectively. In this section, we propose a partial order relation on families of classifiers which is based on the rankings of values of the *test function* in each family relative to that in the optimal family of classifiers (known as Neyman Pearson classifiers):

Definition 2. *Two families of classifiers $\{\bar{H}_t(x)\}$ and $\{H_t(x)\}$ associated with two test functions $\bar{r}(x)$ and $r(x)$ are of the relation $\{\bar{H}_t\} \leq \{H_t\}$, if for any distinct $x_1, x_2 \in \mathbf{R}^d$, we have $\bar{r}(x_1) > \bar{r}_(x_2)$ whenever $r(x_1) \geq r(x_2)$ and $r^o(x_1) > r^o(x_2)$.*

Notice that the minimum of this partial order relation is the optimal family of classifiers, $\{H^o(x)\}$. The following theorem shows that this partial order relation is indeed a partial order relation corresponding to the ROC performances, in which "smaller" families give better ROC curves. The following theorem is regarded as the partial order theorem in the previous section. The proof is in the appendix.

Theorem 3. *If $\{\bar{H}_t\} \leq \{H_t\}$, then the ROC performance associated with the family $\{\bar{H}_t(x)\}$ is better than or at least the same as that associated with the family $\{H_t(x)\}$, i.e. for any fixed false alarm rate, say $P(H_{\hat{t}}(X) = 1|X \sim f(\cdot|2)) = P(\bar{H}_{\bar{t}}(X) = 1|X \sim f(\cdot|2))$ for some $\hat{t}$ and $\bar{t}$, the detection rate associated with the classifier $\bar{H}_{\bar{t}}(x)$ is greater than or equal to that associated with the classifier $H_{\hat{t}}(x)$, $P(\bar{H}_{\bar{t}}(X) = 1|X \sim f(\cdot|1)) \geq P(H_{\hat{t}}(X) = 1|X \sim f(\cdot|1))$.*

Remark 1: Following this theorem, we can define equivalent classes of classifier families, of which the associated test functions are of the same rankings of function values in each equivalent class. In particular, for any test function $r(\cdot)$ with rankings the same as the rankings of $r^o(\cdot)$ (the test function of the optimal family $\{H_t^o\}$), $r(x_1) \leq r(x_2)$ iff $r^o(x_1) \leq r^o(x_2)$, the family of the corresponding classifiers is equivalently optimal (i.e. the corresponding ROC curve is exactly the same as the optimal ROC curve).

Remark 2: If we assume that both false alarm rates $P(\bar{H}_t(X) = 1|X \sim f(\cdot|2))$ and $P(H_t(X) = 1|X \sim f(\cdot|2))$ are continuous functions of t then this theorem implies that the ROC curve associated with the classifier family $\{\bar{H}_{\bar{t}}(x)\}$ is everywhere as least as high as the ROC curve associated with the classifier family $\{H_t(x)\}$.

Next, we study some of the examples discussed in last section to demonstrate the utility of this family-version partial order theorem.

3.1. A training set that contains outliers

Consider a two-class classification problem under the setup in Subsection 2.2.1 where we assume that the underlying density for class i is the linearly perturbed density $\bar{f}(x|i) = \epsilon\hat{f}(x|i) + (1-\epsilon)f(x|i)$ and $\hat{f}(x|i)$ is the outlier density of class i. Consider a family of likelihood ratio classifiers $\bar{H}_t^\epsilon(x)$:

$$\bar{H}_t^\epsilon(x) = \begin{cases} 1 & \text{if } \bar{r}_\epsilon(x) \geq t, \\ 2 & \text{if } \bar{r}_\epsilon(x) < t, \end{cases} \quad \text{where} \quad \bar{r}_\epsilon(x) = \frac{\bar{f}(x|1)}{\bar{f}(x|2)} \text{ is the test function.}$$

Then, does the ROC performance associated with the family $\{\bar{H}_t^\epsilon(x)\}$ always get better or at least the same when we eliminate the outlier weight ϵ? In other words, assuming $\tilde{\epsilon} < \epsilon$, for a fixed false alarm rate, say $P(\bar{H}_t^\epsilon(X) = 1|X \sim f(\cdot|2)) = P(\bar{H}_{\tilde{t}}^{\tilde{\epsilon}}(X) = 1|X \sim f(\cdot|2))$ for some t and $\tilde{t}$, is the detection rate associated with the classifier $\bar{H}_{\tilde{t}}^{\tilde{\epsilon}}(x)$ always greater than or at least equal to that associated with the classifier $\bar{H}_t^\epsilon(x)$, $P(\bar{H}_{\tilde{t}}^{\tilde{\epsilon}}(X) = 1|X \sim f(\cdot|1)) \geq P(\bar{H}_t^\epsilon(X) = 1|X \sim f(\cdot|1))$? Unfortunately, against our intuition, it is not true in general unless we include additional assumptions as in the following two propositions. The proofs can be obtained by simply applying Theorem 3 and are therefore skipped.

Proposition 2. *Assume $\hat{f}(\cdot|2) = f(\cdot|2)$ [or $\hat{f}(\cdot|1) = f(\cdot|1)$]. Let $0 \leq \tilde{\epsilon} < \epsilon \leq 1$. Then for any t and $\tilde{t}$ such that $P(\bar{H}_t^\epsilon(X) = 1|X \sim f(\cdot|2)) = P(\bar{H}_{\tilde{t}}^{\tilde{\epsilon}}(X) = 1|X \sim f(\cdot|2))$, we have $P(\bar{H}_{\tilde{t}}^{\tilde{\epsilon}}(X) = 1|X \sim f(\cdot|1)) \geq P(\bar{H}_t^\epsilon(X) = 1|X \sim f(\cdot|1))$.*

Proposition 3. *Assume $\hat{f}(\cdot|2) = \hat{f}(\cdot|1)$. Let $0 \leq \tilde{\epsilon} < \epsilon \leq 1$. Then for any t and $\tilde{t}$ such that $P(\bar{H}_t^\epsilon(X) = 1|X \sim f(\cdot|2)) = P(\bar{H}_{\tilde{t}}^{\tilde{\epsilon}}(X) = 1|X \sim f(\cdot|2))$, we have $P(\bar{H}_{\tilde{t}}^{\tilde{\epsilon}}(X) = 1|X \sim f(\cdot|1)) \geq P(\bar{H}_t^\epsilon(X) = 1|X \sim f(\cdot|1))$.*

In the first example, we assume $\hat{f}(\cdot|2) = f(\cdot|2)$, which means that there are no outliers in the training set of class 2. This is sometimes feasible in practice. For example, for classification of "heathy patient samples (class 2) versus cancer patient samples (class 1)," the assumption of negligible outliers in heathy patient samples is acceptable given a sufficiently large training set of healthy patients. Thus, the underlying density $\bar{f}(\cdot|2) \approx f(\cdot|2)$ for healthy patients is nearly fixed. The underlying density for cancer patients, $\bar{f}(x|1) = \epsilon\hat{f}(x|1) + (1-\epsilon)f(x|1)$, can be regarded as a learning consequence when a fraction ϵ of the samples in the cancer training set are outliers from distribution $\hat{f}(x|1)$.

In the second example, we assume $\hat{f}(\cdot|1) = \hat{f}(\cdot|2)$. This assumption is sometimes used in robustness analysis (see Chen, Gao, & Ren, 2016; Huber, 1965), where the outliers for different classes or hypotheses are from the same distribution. Thus, the proposition 3.1 can also be applied to study monotone properties of ROC curves in robustness analysis.

3.2. Labeling errors in a training set

Returning to the example of labeling errors in Subsection 2.2.2, consider the 2-class classification problem $(m = 2)$ and consider a family of classifiers of the form

$$H_t^\alpha(x) = \begin{cases} 1 & \text{if } r_\alpha(x) \geq t, \\ 2 & \text{if } r_\alpha(x) < t. \end{cases}$$

where $\alpha = (\alpha_1, \alpha_2)$ and where the test function, $r_\alpha(x) = \frac{\tilde{f}(x|1)}{\tilde{f}(x|2)}$, is the likelihood ratio associated with the underlying densities, $\tilde{f}(x|1)$ and $\tilde{f}(x|2)$, of a training set in which each label i has been independently flipped with probability α_i to one of the other classes. By equation (3), taking $m = 2$, we obtain $r_\alpha(x) = R_\alpha \frac{(1-\alpha_1)p(1)f(x|1)+\alpha_2 p(2)f(x|2)}{(1-\alpha_2)p(2)f(x|2)+\alpha_1 p(1)f(x|1)}$, where $R_\alpha = \frac{(1-\alpha_2)p(2)+\alpha_1 p(1)}{(1-\alpha_1)p(1)+\alpha_2 p(2)}$. Notice that $p(i)$ can be viewed as the underlying probability of class i in training data collection.

To study the ROC performance of $\{H_t^\alpha\}$ in terms of α, according to Theorem 3, we compare the order of the values of $r_\alpha(x_1)$ and $r_\alpha(x_2)$ with the order of $r^o(x_1)$ and $r^o(x_2)$ for any x_1, x_2. We get that $r_\alpha(x_1) \le r_\alpha(x_2)$ iff $(1-\alpha_1-\alpha_2)p(1)p(2)f(x_1|1)f(x_2|2) \le (1-\alpha_1-\alpha_2)p(1)p(2)f(x_2|1)f(x_1|2)$. For $0 \le \alpha_1 + \alpha_2 < 1$,

$$r_\alpha(x_1) \le r_\alpha(x_2) \Longleftrightarrow \frac{f(x_1|1)}{f(x_1|2)} \le \frac{f(x_2|1)}{f(x_2|2)} \Longleftrightarrow r^o(x_1) \le r^o(x_2),$$

and, therefore, according to Remark 1 above, we find, surprisingly, that the ROC curve associated with the test function $r_\alpha(x)$ is the optimal ROC curve. For $1 < \alpha_1 + \alpha_2 \le 2$,

$$r_\alpha(x_1) \le r_\alpha(x_2) \Longleftrightarrow \frac{f(x_1|1)}{f(x_1|2)} \ge \frac{f(x_2|1)}{f(x_2|2)} \Longleftrightarrow r^o(x_1) \ge r^o(x_2),$$

and, therefore, the corresponding ROC curve is symmetric to the optimal ROC curve with respect to the point $(0.5, 0.5)$.

4. NUMERICAL EXPERIMENTS

In Section 2.1, we showed that the CSV error rate is greater than the RCV error rate. To study this variability, we numerically explore the properties of the RCV and CSV methods for three-class classification problems with models learned from gene expression data, similar to the two-class experiment discussed in Chang & Geman (2015). We collected microarray gene expression data of breast cancer patients from (http://watson.compbio.iupui.edu/chirayu/proggene/database/datasources.php), where there are five platforms corresponding to five studies $Z = 1, 2, \cdots, 5$, each of which included expression values for more than 6,000 genes with numbers of patients ranging from one hundred to three hundreds.

For our three-class classification experiments, let $Y = 1, 2$ and 3 denote "relapse free survival", "relapse in three years," and "relapse after three years" respectively. We first quantile-normalized the expression data and then selected differentially expressed genes, say $g_1, g_2, g_3 \cdots, g_p$ by the following procedure on the entire pooled data: First, for each gene, compute the p-value using the Kruskal-Wallis H test. Then, let g_1 be the the gene with smallest p-value. Let g_2 be the gene that has the smallest p-value among those genes whose absolute correlation with g_1 is smaller than 0.25. Let g_3 be the gene that has the smallest p-value among those genes whose absolute correlations with selected genes (i.e., g_1, g_2) are smaller than 0.25. Continue this procedure until the next selected gene has a p-value greater than 0.05, and then we can obtain a set of genes $g_1, g_2, \cdots, g_p$, each of which has p-values smaller that or equal to 0.05 and has absolute correlation less than 0.25 with any other selected gene, where the number of genes $p = 34$. Let $X = (X_1, X_2, \cdots, X_p)$ be predictors denoting the corresponding gene expressions. Then, for each study $Z \in \{1, 2, 3, 4, 5\}$, we use the logistic lasso algorithm from the Glmnet package (Qian et al 2013) to obtain 10 to 12 effective predictors and corresponding nonzero coefficients β_i's for the conditional distribution of Y given X, $f_Z(y|X)$ and we learned a multivariate normal distribution $f_Z(x)$ for X. Therefore, we obtain a joint distribution

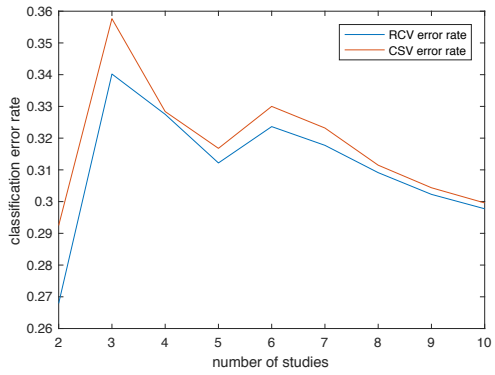

FIGURE 1: **RCV error rates $e_{RCV}(n)$ versus CSV error rates $e_{CSV}(n)$.** The blue and red curves show the RCV error rates $e_{RCV}(n)$ and the CSV error rates $e_{CSV}(n)$ respectively for a particular realization $Z_1, \cdots, Z_{10}$ generated i.i.d. from $U\{1,2,3,4,5\}$, where the number of studies n goes from 2 to 10 and the underlying distributions are $f_{Z_i}(x,y)$, $i = 1, \cdots, 10$.

$f_Z(x,y)$ for each study Z. The algorithms for training and for the following three experiments can be downloaded from https://github.com/Lo-Bin/PartialOrder, and the Glmnet package can be downloaded from https://web.stanford.edu/∼hastie/glmnet_matlab/.

Experiment 1. We generate 10 study variables $Z_1, \cdots, Z_{10}$ uniformly over $\{1,2,3,4,5\}$. For this particular realization $Z_1, \cdots, Z_{10}$, we compute $e_{RCV}(n) = e_{RCV}(Z_1, \cdots, Z_n)$ and $e_{CSV}(n) = e_{CSV}(Z_1, \cdots, Z_n)$ using Monte Carlo integrations with joint distributions $f_{Z_i}(x,y)$, $i = 1, \cdots, n$, where the number of studies $n = 2, \cdots, 10$. Figure 1 shows that the CSV error rate is larger than the RCV error rate for all $n \in \{2, \cdots, 10\}$, which is consistent with Theorem 2. Note that with only two studies Z_1, Z_2 (i.e., $n = 2$), both RCV and CSV error rates are relatively small. However, when adding a third study Z_3 (i.e., $n = 3$), both error rates are largely increased because the classification task for Z_3 is more difficult than that for Z_1, Z_2.

Experiment 2. Unlike the first experiment where we assumed that the distributions $f_{Z_i}(x,y)$'s are known, this experiment explores what happens when the distributions are unknown and have to be estimated. For comparison with the previous experiment, we still use same10 study variables $Z_1, \cdots, Z_{10}$ that are generated in Experiment 1. For each study Z_i, we generate a training set of size 150, $\{(X_{ik}, Y_{ik}) : k = 1, \cdots, 150\}$ from the true distribution $f_{Z_i}(x,y)$, and learn the distribution based on the training set. Next, we construct those Bayes classifiers using the learned distributions, denoted by $\hat{f}_{Z_i}(x,y)$, $i = 1, \cdots, 10$, and compute the RCV and CSV error rates using Monte Carlo integration with true distributions $f_{Z_i}(x,y)$'s as in Experiment 1. Denote the new RCV and CSV error rates by $\hat{e}_{RCV}(n)$ and $\hat{e}_{CSV}(n)$ respectively. The right and left panels of Figure 2 show the results of two runs of the algorithm. Since the size of the training set for each study Z_i is only 150, $\hat{f}_{Z_i}(x,y)$ may not be accurate. Thus, the RCV error rates are larger than in Example 1 because the classifier associated with the RCV approach is based on the mixture of learned distributions [see equation (1)], $\frac{1}{\sum_{i=1}^n \hat{p}_{Z_i}(y)} \sum_{i=1}^n \hat{f}_{Z_i}(x,y)$, which is not accurate. However, the CSV error rates are less affected by the training error. This is because for each $j = 1, \cdots, 10$, when leaving out study Z_j, the classifier associated with CSV is based on the mixture of the learned distributions [see equation (2)], $\frac{1}{\sum_{i\neq j} \hat{p}_{z_i}(y)} \sum_{i\neq j} \hat{f}_{Z_i}(x,y)$, but the error rate is computed based on the true distribution of study Z_j, $f_{Z_j}(x,y)$. Hence, we could sometimes even see the CSV error rate is smaller than the RCV error rate for some n as shown in the right panel of Figure 2, especially when the mixture of learned distributions $\frac{1}{\sum_{i\neq j} \hat{p}_{z_i}(y)} \sum_{i\neq j} \hat{f}_{Z_i}(x,y)$ is more similar to $f_{Z_j}(x,y)$ than the mixture of true distributions

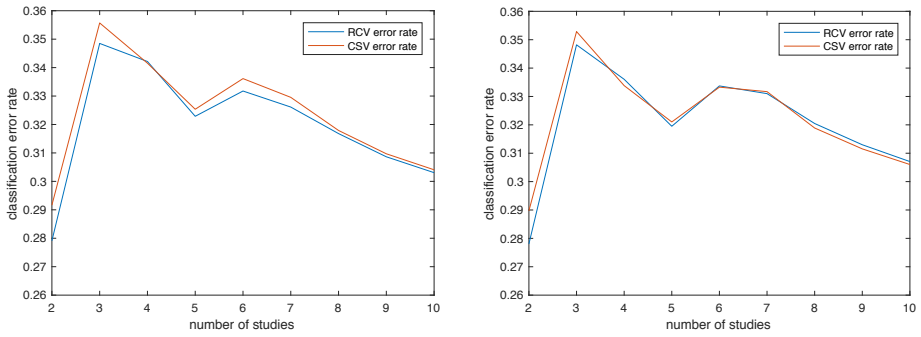

FIGURE 2: **RCV error rates $\hat{e}_{RCV}(n)$ versus CSV error rates $\hat{e}_{CSV}(n)$.** The blue and red curves show the RCV error rates $\hat{e}_{RCV}(n)$ and the CSV error rates $\hat{e}_{CSV}(n)$ respectively for the same realization $Z_1, \cdots, Z_{10}$ as in Experiment 1, where the number of studies n goes from 2 to 10, and the underlying distributions are $f_{Z_i}(x, y)$, $i = 1, \cdots, 10$, but the distributions used to construct the Bayes classifiers are the learned distribution $\hat{f}_{Z_i}(x, y)$, $i = 1, \cdots, 10$. The left and right panels are two runs of the algorithm.

$\frac{1}{\sum_{i\neq j} p_{z_i}(y)} \sum_{i\neq j} f_{Z_i}(x, y)$ to $f_{Z_j}(x, y)$.

Experiment 3. Cross-validation is often used as a tool for model selection to select effective predictors for generalized linear models. For the model used in this section, the multinomial logistic regression $f_Z(y|x)$ for each study $Z \in \{1, 2, 3, 4, 5\}$ has 10 to 12 predictors with nonzero effects. There are total 29 predictors with nonzero effects in at least one of the five studies. In this experiment, we are interested in which of two cross-validation methods, RCV and CSV, better serves model selections. We generate a pooled dataset which includes 150 i.i.d. pairs (X, Y) from each of the five distributions $f_Z(x, y)$, $Z = 1, \cdots, 5$. Then using the logistic lasso, we train on the pooled data to select effective predictors and see if we can discover those 29 predictors. We use the 10-fold cross-validation (default in Glmnet package) and the cross-study validation to determine the penalty coefficient λ in the logistic lasso training. Let λ_{RCV} and λ_{CSV} be the corresponding determined penalty coefficients. Then two sets of effective predictors are selected using λ_{RCV} and λ_{CSV}. Let N_{RCV} and N_{CSV} be the numbers of additional predictors associated with λ_{RCV} and λ_{CSV}, where an "additional predictor" is a selected predictor which is not one of the 29 predictors. Let M_{RCV} and M_{CSV} be the numbers of missed predictors associated with λ_{RCV} and λ_{CSV}, where a "missed predictor" is a non-selected predictor which is one of the 29 predictors.

Next, 1,000 simulations are performed to plot histograms of λ_{rcv},λ_{csv}, (N_{RCV}, N_{CSV}) and (M_{RCV}, M_{CSV}). As shown in Figure 3, λ_{csv} is often larger than λ_{rcv} so that the CSV method tends to select fewer predictors. This makes sense because when the CSV method evaluates the likelihood on the left out study, any predictor selected based on the other four studies becomes a noise term that reduces the likelihood, if the predictor is not an effective predictor of the left out study. Hence, to avoid selecting such a predictor, the CSV method would prefer to use a larger penalty coefficient. As a result, the CSV method selects fewer predictors, which causes a lot more missed predictors, and results in a worse model selection as shown in Figure 4. Notice that the result from the CSV method being so different from that from the RCV method is due to the distribution difference between each left out study and the remaining 4 studies. Therefore, increasing the sample size of each study does not diminish the difference. When increasing the number of studies, m, the difference still remains to some extent because the mixture distribution of any $m - 1$ studies converges to $Ef_Z(x, y)$ (provided that study variables Z_i's are i.i.d. $f(z)$ as assumed in Theorem 2), which could still be different from the distribution of a left out study.

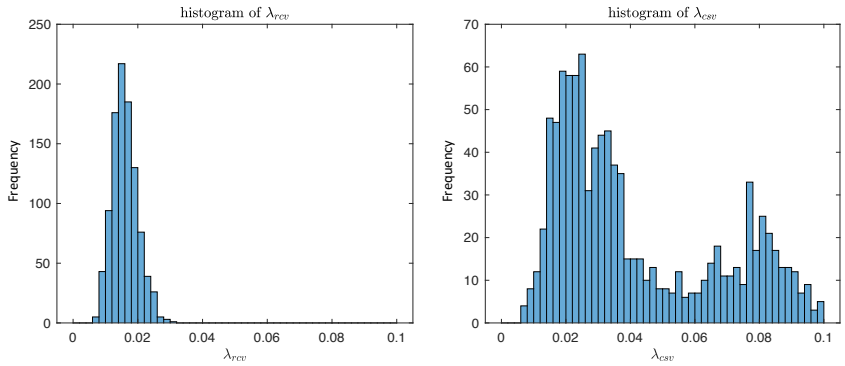

FIGURE 3: **Histograms of λ_{rcv} and λ_{csv}.**

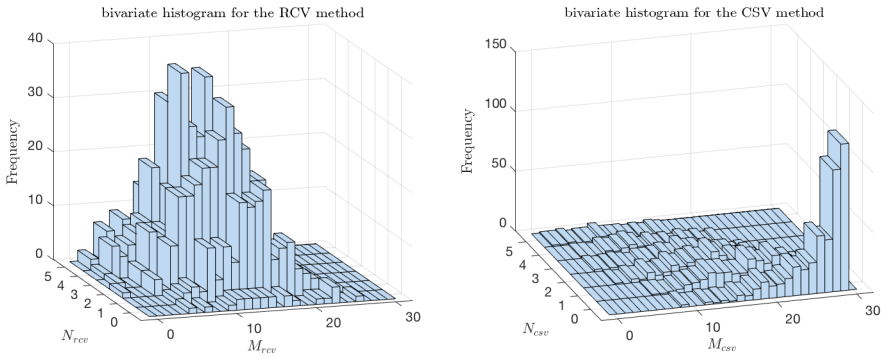

FIGURE 4: **Bivariate Histograms for the RCV and CSV methods.**

5. DISCUSSION

Many challenges arising from this research have remained unexplored. The first partial order relation discussed in this paper is based on the probability of misclassification $\mathrm{Err}(H) = P(H(X) \neq Y)$, which is the risk function associated with loss function $\mathbb{1}_{H(X)\neq Y}$, but many other alternative error measures are available for investigation (e.g. the risk function associated with the hinge loss). In terms of cross-study validation, study effects can be characterized by the gap between the error rate of cross-study validation and the randomized cross-validation. Therefore, with the formulas of these validation error rates given in this paper, further investigation on theoretical studies of quantifying the study effect can be initiated. Finally, in terms of partial order relations between two families of classifiers, if we quantify ROC performance by calculating the area under the ROC curve, can we define a useful partial order relation associated with the area quantification?

As many classification algorithms requires a significant amount of training and classification implementation, several approximation techniques were developed to simplify the computation complexity of training and classification implementation, and reduce computational time and memory usage (e.g., Lee & Huang, 2007; Chang et al., 2013). Therefore, another application of the partial order relations is to analytically compare the classification performances between exact computation and its approximation. Moreover, many articles have demonstrated improvements in classification accuracy experimentally, particularly within the literature on ensemble learning methods such as bagging, boosting, decision trees, stacking, and Bayesian model aver-

aging. With the partial order relations, theoretical investigation of these improvement methods as well as other methods of empirical findings is of particular interest for future research.

ACKNOWLEDGEMENTS

The author thanks Steve MacEachern for his helpful comments and thanks the Ohio supercomputer center for computational support.

BIBLIOGRAPHY

Bootkrajang, J. & Kabán, A. (2012). Label-noise robust logistic regression and its applications. *Machine Learning and Knowledge Discovery in Databases*, 7523,143–158.

Brodley, C.E. & Friedl, M.A. (1999). Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, 11,131–167.

Chang, L.-B., Bai, Z., Huang, S.-Y., & Hwang, C.-R. (2013). Asymptotic error bounds for kernel-based nyström low-rank approximation matrices. *Journal of Multivariate Analysis*, 120, 102–119.

Chang, L.-B. & Geman, D. (2015). Tracking cross-validated estimates of prediction error as studies accumulate. *Journal of the American Statistical Association*, 110, 1239–1247.

Chang, L.-B., Jin, Y., Zhang, W., Borenstein, E., & Geman, S. (2011). Context, computation, and optimal roc performance in hierarchical models. *International Journal of Computer Vision*, 93,117–140.

Chen, M., Gao, C., & Ren, Z. (2016). A general decision theory for Huber's ϵ-contamination model. *Electronic Journal of Statistics*, 10, 3752-3774.

Dudoit, S., Fridlyand, J., & Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97,77–87.

Frenay, B. & Verleysen, M. (2014). Classification in the presence of label noise: a survey. *Neural Networks and Learning Systems*, 25,845–869.

He, X. & Simpson, D.G. & Portnoy, S.L. (1990). Breakdown Robustness of Tests. *Journal of the American Statistical Association*, 85, 446–452.

Huber, P.J. (1965). A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, 36,1753–1758.

Huber, P.J. (1981). Robust Statistics. *John Wiley & Sons, Inc.*

Lee, Y. j. & Huang, S.-Y. (2007). Reduced support vector machines: a statistical theory. *Neural Networks, IEEE Transactions on*, 18,1–13.

Lawrence, N. & Schölkopf, B. (2001). Estimating a kernel fisher discriminant in the presence of label noise. *18th International Conference on Machine Learning*, 306–313.

Lee, Y. & Wang, R. (2015). Does modeling lead to more accurate classification?: A study of relative efficiency in linear classification. *Journal of Multivariate Analysis*, 133, 232–250.

Leung, T. & Song, Y. & Zhang, J. (2011). Handling label noise in video classification via multiple instance learning. *ICCV*, 2056–2063.

Li, C., Wang, J., Wang, L., Hu, L., & Gong, P. (2014). Comparison of classification algorithms and training sample sizes in urban land classification with landsat thematic mapper imagery. *Remote Sensing*, 6, 964–983.

Lim, T.-S., Loh, W.-Y., & Shih, Y.-S. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning*, 40, 203–228.

Ma, S., Sung, J., Magis, A., Wang, Y., Geman, D., & Price, N. (2014). Measuring the effect of inter-study variability on estimating prediction error. *PLOS ONE*, 9, e110840.

Maronna, R., Martin, R.D., & Yohai, V.J. (2006). Robust Statistics: Theory and Methods. *Joh Wiley & Sons, Ltd.*

Nataraja, N., Dhillon, I., & Ravikumar, P. , & Tewari, A. (2013). Learning with Noisy Labels. *Advances in Neural Information Processing Systems 26*, 1196–1204.

Qian, J., Hastie, T., Friedman, J., Tibshirani, R. & Simon, N. (2013). Glmnet for Matlab.

Scott, C. & Blanchard, G. & Handy, G. (2013). Classification with asymmetric label noise: consistency and maximal denoising. *JMLR: Workshop and Conference Proceedings*, 30,1–23.

Wellmann, J., & Gather, U.(1999). A note on contamination models and outliers. *Communications in Statistics - Theory and Methods*, 28, 1793–1802.

Wu, B., Abbott, T., Fishman, D., Mcmurray, W., Mor, G., Stone, K., Ward, D., Williams, K., & Zhao, H. (2003). Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 19,1636–1643.

Zio, M.D., & Guarnera, U.(2013). A contamination model for selective editing. *Journal of Official Statistics*, 29, 539–555.

APPENDIX

Proof of Theorem 1. Let the $\bar{h}_i(x)$ and $h_i(x)$ functions be corresponding rate functions of $\bar{H}$ and H. Let $A_i = \{x : H^o(x) = i\}$, $A_{ij} = \{x \in A_i : H(x) = j\}$ and $A_{ijk} = \{x \in A_{ij} : \bar{H}(x) = k\}$. Thus we have $A_i = \cup_{j=1}^m A_{ij}$ and $A_{ij} = \cup_{k=1}^m A_{ijk}$. With this notation,

$$\mathrm{Err}(H) = \sum_{i=1}^m p(i) \sum_{i'=1}^m \sum_{j\neq i} P_i(A_{i'j}), \text{ and } \mathrm{Err}(\bar{H}) = \sum_{i=1}^m p(i) \sum_{i'=1}^m \sum_{j=1}^m \sum_{k\neq i} P_i(A_{i'jk}).$$

Now consider

$$\begin{aligned}
\mathrm{Err}(H) - \mathrm{Err}(\bar{H}) &= \sum_{i=1}^m p(i) \sum_{i'=1}^m \sum_{j\neq i} P_i(A_{i'j}) - \sum_{i=1}^m p(i) \sum_{i'=1}^m \sum_{j=1}^m \sum_{k\neq i} P_i(A_{i'jk}) \\
&= \sum_{i=1}^m \sum_{i'=1}^m \sum_{j\neq i} \sum_{k=1}^m p(i) P_i(A_{i'jk}) - \sum_{i=1}^m \sum_{i'=1}^m \sum_{j=1}^m \sum_{k\neq i} p(i) P_i(A_{i'jk}) \\
&= \sum_{i=1}^m \sum_{i'=1}^m \sum_{j\neq i} p(i) P_i(A_{i'ji}) - \sum_{i=1}^m \sum_{i'=1}^m \sum_{k\neq i} p(i) P_i(A_{i'ik}).
\end{aligned}$$

Replacing, in the second term, the index i with j and then replacing the index k with i, we have

$$\sum_{i=1}^m \sum_{i'=1}^m \sum_{k\neq i} p(i) P_i(A_{i'ik}) = \sum_{j=1}^m \sum_{i'=1}^m \sum_{i\neq j} p(j) P_j(A_{i'ji}) = \sum_{i=1}^m \sum_{i'=1}^m \sum_{j\neq i} p(j) P_j(A_{i'ji}),$$

where the second equality is obtained by interchanging the order of summation in the triple sum. Therefore,

$$\mathrm{Err}(H) - \mathrm{Err}(\bar{H}) = \sum_{i=1}^m \sum_{i'=1}^m \sum_{j\neq i} \big[p(i) P_i(A_{i'ji}) - p(j) P_j(A_{i'ji})\big]$$

if $x \in A_{i'ji}$, $H(x) = j$ and $\bar{H}(x) = i$ so that we have $h_j(x) \geq h_i(x)$ and $\bar{h}_j(x) \leq \bar{h}_i(x)$. Because $\bar{H} \leq H$, we must have $h_j^o(x) \leq h_i^o(x)$ [i.e. $p(j)f(x|j) \leq p(i)f(x|j)$] if $h_j(x) \geq h_i(x)$ and $\bar{h}_j(x) \leq \bar{h}_i(x)$. Therefore,

$$p(i) P_i(A_{i'ji}) - p(j) P_j(A_{i'ji}) = \int_{A_{i'ji}} \big[p(i) f(x|i) - p(j) f(x|j)\big] dx \geq 0.$$

Hence, $\mathrm{Err}(H) - \mathrm{Err}(\bar{H}) \geq 0$, and the proof is completed. ■

Proof of Theorem 2. For each $j \in \{1, 2, \cdots, m\}$, consider the following two classifiers: $H_j(x) = \arg\max_{t\in\{1,2,\cdots,m\}} \sum_{i\neq j} p_{z_i}(t) f_{z_i}(x|t)$ and $\bar{H}(x) = \arg\max_{t\in\{1,2,\cdots,m\}} \sum_{i=1}^{n} p_{z_i}(t) f_{z_i}(x|t)$. Regarding $p_{z_j}(k)$ and $f_{z_j}(x|k)$ as prior probability $p(k)$ and conditional density $f(x|k)$ defined in the beginning of Section 2 for $k = 1, \cdots, m$, we can get that if $\sum_{i\neq j} p_{z_i}(t_1) f_{z_i}(x|t_1) \geq \sum_{i\neq j} p_{z_i}(t_2) f_{z_i}(x|t_2)$ and $h^o_{t_1}(x) = p_{z_j}(t_1) f_{z_j}(x|t_1) > h^o_{t_2}(x) = p_{z_j}(t_2) f_{z_j}(x|t_2)$, then we have $\sum_{i=1}^{n} p_{z_i}(t_1) f_{z_i}(x|t_1) > \sum_{i=1}^{n} p_{z_i}(t_2) f_{z_i}(x|t_2)$. Therefore, $\bar{H}(x) \leq H_j(x)$ and according to the partial order theorem, $\sum_{k=1}^{m} p_{z_j}(k) \cdot P\left(H_j(X) \neq k |\, X \sim f_{z_j}(\cdot|k)\right) \geq \sum_{k=1}^{m} p_{z_j}(k) \cdot P\left(\bar{H}(X) \neq k \big|\, X \sim f_{z_j}(\cdot|k)\right)$. This implies

$$
\begin{aligned}
e_{CSV}(z_1, ..., z_n) &= \frac{1}{n} \sum_{j=1}^{n} \sum_{k=1}^{m} p_{z_j}(k) \cdot P\left(H_j(X) \neq k |\, X \sim f_{z_j}(\cdot|k)\right) \\
&\geq \frac{1}{n} \sum_{j=1}^{n} \sum_{k=1}^{m} p_{z_j}(k) \cdot P\left(\bar{H}(X) \neq k \big|\, X \sim f_{z_j}(\cdot|k)\right) \qquad (1)
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
& P\left(\bar{H}(X) \neq k \middle| X \sim \frac{1}{\sum_{i=1}^{n} p_{z_i}(k)} \sum_{i=1}^{n} p_{z_i}(k) f_{z_i}(\cdot|k)\right) \\
&= \int_{\{x:\bar{H}(x)\neq k\}} \frac{1}{\sum_{i=1}^{n} p_{z_i}(k)} \sum_{i=1}^{n} p_{z_i}(k) f_{z_i}(x|k) dx \\
&= \frac{1}{\sum_{i=1}^{n} p_{z_i}(k)} \sum_{i=1}^{n} p_{z_i}(k) \int_{\{x:\bar{H}(x)\neq k\}} f_{z_i}(x|k) dx \\
&= \frac{1}{\sum_{i=1}^{n} p_{z_i}(k)} \sum_{i=1}^{n} p_{z_i}(k) P\left(\bar{H}(X) \neq k | X \sim f_{z_j}(\cdot|k)\right).
\end{aligned}
$$

Thus we can rewrite equation (1) and obtain

$$
\begin{aligned}
e_{CSV}(z_1, ..., z_n) &\geq \sum_{k=1}^{m} \frac{\sum_{i=1}^{n} p_{z_i}(k)}{n} \cdot P\left(\bar{H}(X) \neq k \middle| X \sim f_{z_{1:n}}(x|k)\right) \\
&= e_{RCV}(z_1, ..., z_n).
\end{aligned}
$$

■

Proof of Theorem 3. Let $r(x)$ and $\hat{r}(x)$ be the test functions associated with $H(x)$ and $\hat{H}(x)$, and assume that $\hat{t}$ and $\bar{t}$ are thresholds such that $H_{\hat{t}}$ and $\bar{H}_{\bar{t}}$ have the same false alarm rate, $P(H_{\hat{t}}(X) = 1|X \sim f(\cdot|2)) = P(\bar{H}_{\bar{t}}(X) = 1|X \sim f(\cdot|2))$. Define $A = \{x : H_{\hat{t}}(x) = 1\} = \{x : r(x) \geq \hat{t}\}$, $B = \{x : \bar{H}_{\bar{t}}(x) = 1\} = \{x : \bar{r}(x) \geq \bar{t}\}$, $D = A \cap B$, $\tilde{A} = A \setminus D$, and $\tilde{B} = B \setminus D$. By the assumption of $\hat{t}$ and $\bar{t}$, we have $\int_A f(x|2)dx = \int_B f(x|2)dx$ and thus

$$
\int_{\tilde{A}} f(x|2)dx = \int_{\tilde{B}} f(x|2)dx. \qquad (2)
$$

Similarly, to prove $P(\bar{H}_{\bar{t}}(X) = 1|X \sim f(\cdot|1)) \geq P(H_{\hat{t}}(X) = 1|X \sim f(\cdot|1))$, it is sufficient to show that $\int_{\tilde{A}} f(x|1)dx \leq \int_{\tilde{B}} f(x|1)dx$. Since $\tilde{A} = \{x : \bar{r}(x) < \bar{t}\} \cap \{x : r(x) \geq \hat{t}\}$ and $\tilde{B} = \{x : \bar{r}(x) \geq \bar{t}\} \cap \{x : r(x) < \hat{t}\}$, for any $x_1 \in \tilde{A}, x_2 \in \tilde{B}$, we have $r(x_1) \geq \hat{t} > r(x_2)$ and $\bar{r}(x_1) < \bar{t} \leq \bar{r}(x_2)$ so by the definition of partial order relation $\{\bar{H}_t\} \leq \{H_t\}$, we get $\frac{f(x_1|1)}{f(x_1|2)} \leq \frac{f(x_2|1)}{f(x_2|2)}$. Therefore, there exists $t^* \geq 0$ such that, for any $x_1 \in \tilde{A}, x_2 \in \tilde{B}$, we have $\frac{f(x_1|1)}{f(x_1|2)} \leq t^* \leq \frac{f(x_2|1)}{f(x_2|2)}$. Hence, we get $f(x|1) \leq t^* f(x|2)$ for any $x \in \tilde{A}$, which implies $\int_{\tilde{A}} f(x|1)dx \leq \int_{\tilde{A}} t^* f(x|2)dx$, and $t^* f(x|2) \leq f(x|1)$ for any $x \in \tilde{B}$, which implies $\int_{\tilde{B}} t^* f(x|2)dx \leq \int_{\tilde{B}} f(x|1)dx$. By equation (2), together with the above two implied inequalities, we complete the proof. ■

Received 9 July 2009
Accepted 8 July 2010