Semi-Supervised Learning of Classifiers from a Statistical Perspective: A Brief Review

Daniel Ahfock, Geoffrey J. McLachlan*

Department of Mathematics, University of Queensland, St. Lucia, Brisbane, Australia

ARTICLE INFO

ABSTRACT

There has been increasing attention to semi-supervised learning (SSL) approaches in machine learning to forming a classifier in situations where the training data for a classifier consists of a limited number of classified observations but a much larger number of unclassified observations. This is because the procurement of classified data can be quite costly due to high acquisition costs and subsequent financial, time, and ethical issues that can arise in attempts to provide the true class labels for the unclassified data that have been acquired. A review is provided of statistical SSL approaches to this problem, focussing on the recent result that a classifier formed from a partially classified sample can actually have smaller expected error rate than that if the sample were completely classified. This rather paradoxical outcome is able to be achieved by introducing a framework with a missingness mechanism for the missing labels of the unclassified observations. It is most relevant in commonly occurring situations in practice, where the unclassified data occur primarily in regions of relatively high entropy in the feature space thereby making it difficult for their class labels to be easily obtained.

1. Introduction

Due to the scarcity and often high acquisition cost of labelled data, machine learning methods that make effective use of large quantities of unlabelled data are being increasingly used. One such method is semi-supervised learning (SSL) where, in addition to labelled data, possibly large numbers of unlabelled observations are available at the time of the construction of the classification rule (classifier) to be used. Not surprisingly, semi-supervised learning approaches have been gaining much attention in both the application oriented and the theoretical machine learning communities.

However, theoretical analysis of SSL has so far been scarce. But recently, Ahfock and McLachlan (2020) provided an asymptotic basis on how to increase in certain situations the accuracy of the commonly used linear discriminant function formed from a partially classified sample as in SSL. The increase in accuracy can be of sufficient magnitude for this SSL-based classifier to have smaller error rate than that if it were formed from a completely classified sample. This apparent paradox can be explained by the fact that by introducing the concept of missingness for the unobserved (missing) labels, the information in the missing labels has the potential to compensate for the loss of information from not knowing the missing labels to the extent where this rule can have lower error rate than that if the sample were completely classified

---

* Corresponding author.
  E-mail addresses: g.mclachlan@uq.edu.au, gjm@maths.uq.edu.au (G.J. McLachlan).

in situations where the missingness-mechanism of the class labels is non-ignorable in the pioneering framework of Rubin (1976) for missingness in incomplete-data analysis.

In this paper, we are to focus on SSL from a statistical perspective reviewing, in particular, recent results. Also, we provide some new results in Lemma 1 and Theorems 3 and 4 on the asymptotic relative efficiency of the classifier formed ignoring the information that is available in situations where the labels of the unclassified features are missing at random but non-ignorable.

2. Brief Overview of SSL Approaches

Methods for semi-supervised learning (SSL) typically involve one or more of the following assumptions.

(i) **Smoothness assumption:** If two feature vectors $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ in a high-density region are close, then so should be their corresponding outputs (class labels) z_1 and z_2.
(ii) **Cluster assumption:** If points are in the same cluster, they are likely to be of the same class.
(iii) **Low-density separation assumption:** The decision boundary should lie in a low-density region.
(iv) **Manifold assumption:** High-dimensional data lie (roughly) on a low-dimensional manifold.

These modelling assumptions create a paradigm where unlabelled data can be useful for model estimation (Chapelle et al., 2010). There has been a recent revival of interest in semi-supervised learning in the machine learning community due to impressive empirical progress on benchmark image and text classification datasets; for example, Tarvainen and Valpola (2017), Laine and Aila (2017), Miyato et al. (2019), Berthelot et al. (2019), Xie et al. (2019), Berthelot et al. (2020), Sohn et al. (2020), and Wei et al. (2021).

From a statistical perspective, the theoretical understanding of these modern semi-supervised learning algorithms is still limited. One of the most intuitive approaches to semi-supervised learning is self-training, an iterative method for learning with alternating steps between generating pseudo-labels for the unlabelled observations and then training a classifier using both the labelled and pseudo-labelled data. As the algorithm progresses both the classification rule and the pseudo-labelling rule are updated. The underlying intuition is very similar to the underlying principles of the expectation–maximization (EM) algorithm of Dempster et al. (1977), in that one iterates between generating pseudo-labels to form the complete data and then estimating a classification model on the basis of the complete data. In early approaches, the same model is used for the classification rule and the generation of pseudo-labels. However, this has been found to be inefficient in some cases (Chapelle et al., 2010). Modern approaches often use separate models for the classifier and the pseudo-labeller, although the training of each can be coupled (Chen et al., 2020b; Tarvainen and Valpola, 2017; Wang et al., 2020). Self-training can also be used to check that estimated error rates for a classification rule are compatible with certain targets, such as homogeneous false positive rates across demographic groups. This is particularly relevant when the fairness or algorithmic bias is a concern (Ji et al., 2020). Unlabelled data are also used to construct data-dependent regularisation terms. Two of the most important types are entropy regularisation and consistency regularisation. Entropy regularisation enforces the belief that the classifier should make confident predictions on the unlabelled data (Grandvalet and Bengio, 2005). Entropy regularisation is connected to the low-density separation assumption, as models with the decision boundary in low-density regions will generate predictions with smaller expected entropy compared to models which place the decision boundary through a high-density region. Consistency regularisation encourages the model to make similar predictions for unlabelled observations that are close together in feature-space (Belkin et al., 2006). This is closely connected to the semi-supervised smoothness assumption.

Data-augmentation is another important technique that is often used to improve data-dependent regularisation. Additional unlabelled observations are generated by perturbing the original set of unlabelled observations by adding random noise or transformations such as rotations or translations. Data augmentation is typically coupled with consistency regularisation so that similar predictions are encouraged on the original instances and the augmented versions (Berthelot et al., 2019; Nair et al., 2019; Wei et al., 2021). The combined use of small local alterations and more aggressive global changes has been found to be an effective strategy (Sohn et al., 2020). The generating process for the augmentations can be related to the manifold assumption, as ideally the transformed data-points help to learn the natural low-dimensional manifold more accurately. This is particularly relevant when working with images (Horlava et al., 2020). Data augmentation has been demonstrated to improve the effectiveness of consistency regularisation and entropy regularisation (Berthelot et al., 2019; Nair et al., 2019; Sohn et al., 2020; Wei et al., 2021).

There are many open research questions regarding semi-supervised learning. A fundamental issue is identifying the mechanisms responsible for the success of semi-supervised learning. For generative classifiers that model the joint distribution of the features and the labels (McLachlan, 1992), it is possible to compute the Fisher information in unclassified observations and the benefits of using unlabelled observations are well understood. For discriminative classifiers that model the conditional distribution of the labels given the features (Ng and Jordan, 2002), quantifying the available information in the unlabelled data is a very difficult problem (Lafferty and Wasserman, 2007; Singh et al., 2008; Göpfert et al., 2019). There are some theoretical results that show unlabelled data can improve estimation efficiency (Sokolovska et al., 2008; Yuval and Rosset, 2020). However, they do not fully explain the success of data-dependent regularisation and data augmentation. Recent theoretical work has highlighted the ability of data augmentation to encode prior knowledge about the classification problem, by enforcing invariance patterns that should hold under the generative model

(Chen et al., 2020a; Decoste and Schökopf, 2002). The exploitation of task invariances is a general strategy for variance reduction in machine learning (Decoste and Schökopf, 2002; Teo et al., 2008).

3. Optimal classifier specified by the Bayes' rule

In the sequel, we focus on statistical approaches to SSL using parametric approaches to the estimation of the (optimal) Bayes' rule from a partially classified sample.

We let $\boldsymbol{y}$ be a p-dimensional vector of features on an entity to be assigned to one of g predefined classes $C_1, \ldots, C_g$. The random variable $\boldsymbol{Y}$ corresponding to the realisation $\boldsymbol{y}$ is assumed to have density $f_i(\boldsymbol{y}; \boldsymbol{\omega}_i)$ known up to a vector $\boldsymbol{\omega}_i$ of unknown parameters in Class C_i $(i = 1, \ldots, g)$. The optimal (Bayes') rule of allocation $R(\boldsymbol{y}; \boldsymbol{\theta})$ assigns an entity with feature vector $\boldsymbol{y}$ to Class C_k (that is, $R(\boldsymbol{y}; \boldsymbol{\theta}) = k$) if

$$k = \arg \max_i \ \tau_i(\boldsymbol{y}; \boldsymbol{\theta}), \tag{1}$$

where

$$\tau_i(\boldsymbol{y}; \boldsymbol{\theta}) = \pi_i f_i(\boldsymbol{y}; \boldsymbol{\omega}_i) / \sum_{h=1}^{g} \pi_h f_h(\boldsymbol{y}; \boldsymbol{\omega}_h) \tag{2}$$

is the posterior probability that the entity belongs to Class C_i given $\boldsymbol{Y} = \boldsymbol{y}$ and π_i is the prior probability that the entity belongs to C_i $(i = 1, \ldots, g)$; $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_{g-1}, \boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_g^T)^T$ is the vector of unknown parameters. The superscript T denotes vector transpose.

In order to estimate $\boldsymbol{\theta}$, it is customary in practice to have available a training sample of size n. We let $\boldsymbol{x}_{\text{CC}} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_n^T)^T$ contain n independent realisations of $X = (\boldsymbol{Y}^T, Z)^T$ as the completely classified training data, where Z denotes the class membership of $\boldsymbol{Y}$, being equal to i if $\boldsymbol{Y}$ belongs to class C_i $(i = 1, \ldots, g)$, and zero otherwise, and where $\boldsymbol{x}_j = (\boldsymbol{y}_j^T, z_j)^T$. For a partially classified training sample $\boldsymbol{x}_{\text{PC}}$ in SSL, we introduce the missing-label indicator m_j which equals 1 if z_j is missing and 0 if it is available $(j = 1, \ldots, n)$. Thus $\boldsymbol{x}_{\text{PC}}$ consists of those observations $\boldsymbol{x}_j$ in $\boldsymbol{x}_{\text{CC}}$ with $m_j = 0$, but only the feature vector $\boldsymbol{y}_j$ in $\boldsymbol{x}_{\text{CC}}$ if $m_j = 1$ (that is, the label z_j is missing).

Considerable simplification is possible under the two-class (homoscedastic) normal model,

$$\boldsymbol{Y} \mid Z = i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}) \quad \text{in } C_i \quad \text{with prob. } \pi_i \quad (i = 1, 2). \tag{3}$$

Under (3), the Bayes' rule reduces to depending on just the $(p + 1)$-dimensional vector of discriminant function coefficients $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^T)^T$, since $R(\boldsymbol{y}; \boldsymbol{\theta})$ is 1 or 2, according as

$$d(\boldsymbol{y}; \boldsymbol{\beta}) = \beta_0 + \boldsymbol{\beta}_1^T \boldsymbol{y} \tag{4}$$

is greater or less than zero, where

$$\beta_0 = \log(\pi_1/\pi_2) - \tfrac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \tag{5}$$

$$\boldsymbol{\beta}_1 = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{6}$$

We can reparameterize the two-class normal model (3) by taking $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\beta}^T)^T$, where $\boldsymbol{\theta}_1$ contains the elements of $\boldsymbol{\mu} = \pi_1 \boldsymbol{\mu}_1 + \pi_2 \boldsymbol{\mu}_2$ and the distinct elements of

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma} + \pi_1 \pi_2 (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T.$$

It has the convenient canonical form,

$$\boldsymbol{\Sigma} = \boldsymbol{I}_p, \ \boldsymbol{\mu}_1 = (\Delta, 0, \ldots, 0)^T, \ \boldsymbol{\mu}_2 = (0, \ldots, 0)^T, \tag{7}$$

where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix and

$$\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

is the squared Mahalanobis distance between the two classes.

4. History of SSL in Statistics

In the statistical community, the so-called self-training approach to SSL in a sense was suggested by C.A.B. Smith (FRS), the distinguished British statistician and geneticist, who held the Weldon Chair of Biometry at University College London. In his discussion of the paper titled *Allocation Rules and Their Error Rates* read to the Royal Statistical Society by Hills (1966), Smith (1966) suggested that in the case of a completely unclassified sample which exhibits bimodality on some feature, a classifier be formed from the unclassified observations on the feature as follows: "One then arbitrarily divides them at the antimode,.... On the basis of this division, we calculate a suitable allocation rule; and, by using this allocation rule, get an

improved division, and so on. As far as I know, there is no theoretical research into the effect of 'lifting oneself by one's own bootstraps' in this way."

It subsequently led McLachlan (1975) to consider this iterative approach as suggested by Smith (1966) on a theoretical basis. The parameters are estimated iteratively by treating the labels of the unclassified features as unknown parameters to be estimated along with the parameters of the allocation rule. That is, it uses the so-called classification maximum likelihood (CML) approach as considered by Hartley and Rao (1968), among others; see Section 1.12 of McLachlan and Basford (1988). The CML approach gives an inconsistent estimate of θ except in special cases like $\pi_1 = \pi_2$. The CML approach can be viewed as applying the subsequent EM algorithm of Dempster et al. (1977) with the following modification (McLachlan, 1982). Namely, the E-step is executed using outright (hard) rather than fractional (soft) assignment of each unclassified feature to a component of the mixture as with the standard application of the EM algorithm.

In order to make the problem analytically tractable for the calculation of the expected error rate of the estimated Bayes' rule, McLachlan (1975) assumed that there were also a limited number n_{ic} of classified features available from C_i in addition to the number of $n_u = n - n_c$ unclassified features, where n denotes the total size of the now partially classified sample and $n_c = n_{1c} + n_{2c}$. We let $\hat{\boldsymbol{\beta}}_{PC}^{(k)}$ denote the estimator after the kth iteration of the vector $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^T)^T$ of discriminant function coefficients obtained by the CML approach applied to the partially classified sample $\boldsymbol{x}_{PC}$. The estimated Bayes' rule using $\hat{\boldsymbol{\beta}}_{PC}^{(k)}$ for $\boldsymbol{\beta}$ in the Bayes' rule $R(\boldsymbol{y}; \boldsymbol{\beta})$ is denoted by $\hat{R}_{PC}^{(k)}$.

McLachlan (1975) showed in the case of known equal prior probabilities that the overall expected error rate of this classifier $\hat{R}_{PC}^{(k)}$ after the kth iteration is given, as $n_u \to \infty$, by

$$\Phi(-\tfrac{1}{2}\Delta) + \{\phi(\tfrac{1}{2}\Delta)/4\}\, a_1^{(k)} + O(n_c^{-2}), \tag{8}$$

where

$$a_1^{(k)} = h_1^{2k}\frac{\Delta}{4} + h_2^{2k}\frac{p-1}{\Delta}\left(\frac{1}{n_{1c}} + \frac{1}{n_{2c}}\right) + h_2^{2k}\frac{(p-1)\Delta}{n_c - 2},$$

$$h_1 = \phi(\tfrac{1}{2})[4\phi(\tfrac{1}{2}) + \Delta\{1 - 2\Phi(-\tfrac{1}{2})\}],$$
$$h_2 = \{\phi(\tfrac{1}{2})\}^2\,(4 + \Delta^2)/h_1,$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal density and (cumulative) distribution, respectively, function.

As it can be shown that both $|h_1|$ and $|h_2|$ are always less than one, it follows from (8) that the expected error rate of $\hat{R}_{PC}^{(k)}$ decreases after each iteration and converges to the optimal error rate $\Phi(-\tfrac{1}{2}\Delta)$, as $k \to \infty$.

McLachlan (1977) considered the above approach to estimating the linear discriminant function in the case where the number of unclassified feature vectors was small rather than very large. The idea was to weight the estimates of the class means based on the unclassified data with those based on the classified data. That is, the class estimates based on the classified data were distinguished in importance from those based on the unclassified data by the introduction of a weighting factor. But it is preferable to weight each feature vector individually as to whether it is unclassified or not, which is able to be achieved with the application of the EM algorithm which appeared also in 1977.

More recently, a process called termed fractional supervised classification (FSC) was proposed whereby the likelihoods L_C and L_{UC} were weighted by a single weighting factor (Vbrik and McNicholas, 2015; Gallaugher and McNicholas, 2019). The reader is referred to McLachlan and Ahfock (2021) for some simulations on the FSC approach where the estimated rule is applied to data subsequent to the partially classified training sample.

5. Estimation of the Bayes' Classifier

The construction of the parametric version of the optimal (Bayes') classifier from partially classified data can be undertaken by maximum likelihood (ML) estimation of $\boldsymbol{\theta}$ implemented via the EM algorithm of Dempster et al. (1977); see also McLachlan and Krishnan (2008). We let

$$\log L_C(\boldsymbol{\theta}) = \sum_{j=1}^{n}(1 - m_j)\sum_{i=1}^{g} z_{ij}\log\{\pi_i f_i(\boldsymbol{y}_j; \boldsymbol{\omega}_i)\}, \tag{9}$$

$$\log L_{UC}(\boldsymbol{\theta}) = \sum_{j=1}^{n} m_j \log\sum_{i=1}^{g}\pi_i f_i(\boldsymbol{y}_j; \boldsymbol{\omega}_i), \tag{10}$$

$$\log L_{PC}^{(ig)}(\boldsymbol{\theta}) = \log L_C(\boldsymbol{\theta}) + \log L_{UC}(\boldsymbol{\theta}), \tag{11}$$

where in (9), $z_{ij} = 1$ if $z_j = i$ and is zero otherwise.

In situations where one proceeds by ignoring the "missingness" of the class labels, $L_C(\boldsymbol{\theta})$ and $L_{UC}(\boldsymbol{\theta})$ denote the likelihood function formed from the classified data and the unclassified data, respectively, and $L_{PC}^{(ig)}(\boldsymbol{\theta})$ is the likelihood function formed

from the partially classified sample $\boldsymbol{x}_{PC}$, ignoring the missing-data mechanism for the labels. The log of the likelihood $L_{CC}(\boldsymbol{\theta})$ for the completely classified sample $\boldsymbol{x}_{CC}$ is given by (9) with all $m_j = 0$.

In later sections, we are to consider situations where there is a link between the pattern of missing labels and class uncertainty. In such situations, the likelihood function (11) is referred to as the likelihood that ignores the missing-label mechanism.

We let $\hat{\boldsymbol{\theta}}_{CC}$ and $\hat{\boldsymbol{\theta}}_{PC}^{(ig)}$ be the estimate of $\boldsymbol{\theta}$ formed by consideration of $L_{CC}(\boldsymbol{\theta})$ and $L_{PC}^{(ig)}(\boldsymbol{\theta})$, respectively. Also, we let $R(\boldsymbol{y}; \hat{\boldsymbol{\theta}}_{CC})$ and $R(\boldsymbol{y}; \hat{\boldsymbol{\theta}}_{PC}^{(ig)})$ denote the estimated Bayes' rule obtained by plugging in $\hat{\boldsymbol{\theta}}_{CC}$ and $\hat{\boldsymbol{\theta}}_{PC}^{(ig)}$, respectively, for $\boldsymbol{\theta}$ in $R(\boldsymbol{y}; \boldsymbol{\theta})$. The overall conditional error rate of the rule $R(\boldsymbol{y}; \hat{\boldsymbol{\theta}}_{CC})$ is defined by

$$\text{err}(\hat{\boldsymbol{\theta}}_{CC}; \boldsymbol{\theta}) = 1 - \sum_{i=1}^{g} \pi_i \, \text{pr}\{R(\boldsymbol{Y}; \hat{\boldsymbol{\theta}}_{CC}) = i \mid \hat{\boldsymbol{\theta}}_{CC}, Z = i\}. \tag{12}$$

The corresponding conditional error rate $\text{err}(\hat{\boldsymbol{\theta}}_{PC}^{(ig)}; \boldsymbol{\theta})$ of the rule $R(\boldsymbol{y}; \hat{\boldsymbol{\theta}}_{PC}^{(ig)})$ is defined by replacing $\hat{\boldsymbol{\theta}}_{CC}$ with $\hat{\boldsymbol{\theta}}_{PC}^{(ig)}$ in (12). The optimal error rate $\text{err}(\boldsymbol{\theta})$ is defined by replacing $\hat{\boldsymbol{\theta}}_{CC}$ by $\boldsymbol{\theta}$ in (12).

6. ARE of $\hat{R}_{PC}^{(ig)}$ in case of MCAR labels

The concepts of *missing at random* (MAR) and its more restrictive version *missing completely at random* (MCAR) are key concepts in the missing-data framework of Rubin (1976) for missingness in incomplete-data analysis. The reader is referred to the clarification note of of Mealli and Rubin (2015) for precise definitions of MAR and MCAR and their corresponding more restrictive versions, *missing always at random* and *missing always completely at random*, respectively. Their paper also gives the weakest sufficient conditions under which ignoring the missingness mechanism that creates the missing data always leads to appropriate inferences about $\boldsymbol{\theta}$ under frequentist, direct likelihood, or Bayesian modes of inferences.

In this section, it is assumed that the unclassified feature vectors have labels that are MCAR. The condition of the missing data to be *missing (always) completely at random* when coupled with exchangeability is a statement about conditional independence in that it implies that the vector of missing-label indicators M_j is independent of the data (Mealli and Rubin, 2015). Here this implies that M_j is independent of $\boldsymbol{Y}_j$ and Z_j ($j = 1, \ldots, n$).

The relative value of the partially classified sample to the completely classified sample can be measured by comparing the expected excess error rate of a classifier trained on each of the respective samples $\boldsymbol{x}_{PC}$ and $\boldsymbol{x}_{CC}$ (Ganesalingam and McLachlan, 1978; O'Neill, 1978; Castelli and Cover, 1996).

As seen in Section 3, in the case of the two-class homoscedastic normal model (3), the Bayes' rule depends only on the vector of discriminant coefficients $\boldsymbol{\beta}$ as defined by (5) and (6). We let $\hat{R}_{PC}^{(ig)} = R(\boldsymbol{y}; \hat{\boldsymbol{\beta}}_{PC}^{(ig)})$ and $\hat{R}_{CC} = R(\boldsymbol{y}; \hat{\boldsymbol{\beta}}_{CC})$ be the estimated Bayes' rules obtained by plugging the estimates $\hat{\boldsymbol{\beta}}_{PC}^{(ig)}$ and $\hat{\boldsymbol{\beta}}_{CC}$, respectively, for $\boldsymbol{\beta}$ into (4).

The relative efficiency of $\hat{R}_{PC}^{(ig)}$ based on the partially classified sample $\boldsymbol{x}_{PC}$ compared to the CSL rule $\hat{R}_{CC}$ based on the completely classified sample $\boldsymbol{x}_{CC}$ is defined as

$$\text{RE}(\hat{R}_{PC}^{(ig)}) = \frac{E\{\text{err}(\hat{\boldsymbol{\beta}}_{CC}; \boldsymbol{\beta})\} - \text{err}(\boldsymbol{\beta})}{E\{\text{err}(\hat{\boldsymbol{\beta}}_{PC}^{(ig)}; \boldsymbol{\beta})\} - \text{err}(\boldsymbol{\beta})}. \tag{13}$$

The asymptotic relative efficiency (ARE) of the rule $\hat{R}_{PC}^{(full)}$ compared to the rule $\hat{R}_{CC}$ is defined as

$$\text{ARE}(\hat{R}_{PC}^{(ig)}) = \frac{AE\{\text{err}(\hat{\boldsymbol{\beta}}_{CC}; \boldsymbol{\beta})\} - \text{err}(\boldsymbol{\beta})}{AE\{\text{err}(\hat{\boldsymbol{\beta}}_{PC}^{(ig)}; \boldsymbol{\beta})\} - \text{err}(\boldsymbol{\beta})}, \tag{14}$$

where $AE\{\text{err}(\hat{\boldsymbol{\beta}}_{PC}; \boldsymbol{\beta})\}$ and $AE\{\text{err}(\hat{\boldsymbol{\beta}}_{CC}; \boldsymbol{\beta})\}$ denote the expansion of the expected error rate $E\{\text{err}(\hat{\boldsymbol{\beta}}_{PC}; \boldsymbol{\beta})\}$ and of $E\{\text{err}(\hat{\boldsymbol{\beta}}_{CC}; \boldsymbol{\beta})\}$, respectively, up to and including terms of the first order with respect to $1/n$.

Under the assumption that the class labels are *missing always completely at random* (that is, the missingness of the labels does not depend on the data), Ganesalingam and McLachlan (1978) derived the ARE of $\hat{R}_{PC}^{(ig)}$ compared to $\hat{R}_{CC}$ in the case of a completely unclassified sample ($\gamma = 1$) for univariate features ($p = 1$). Their results are listed in Table 1 for $\Delta = 1, 2, 3$, and 4.

O'Neill (1978) made use of a result of Efron (1975) on the asymptotic relative efficiency of logistic regression to extend the results of Ganesalingam and McLachlan (1978) to multivariate features and for arbitrary γ. His results showed that this ARE was not sensitive to the values of p and does not vary with p for equal class prior probabilities. It can be seen from Table 1 that the ARE of $\hat{R}_{PC}^{(ig)}$ for a totally unclassified sample is low, particularly for classes weakly separated as represented by $\Delta = 1$ in Table 1.

In order to obtain the ARE of $\hat{R}_{PC}^{(ig)}$ compared to $\hat{R}_{CC}$, O'Neill (1978) derived the information matrix $\boldsymbol{I}_{PC}(\boldsymbol{\beta})$ about the vector $\boldsymbol{\beta}$ of discriminant function coefficients in the context of the two-class normal discrimination problem with parameter vector

Table 1

Asymptotic relative efficiency of $\hat{R}_{PC}^{(ig)}$ compared to $\hat{R}_{CC}$

π_1	$\Delta = 1$	$\Delta = 2$	$\Delta = 3$	$\Delta = 4$
0.1	0.0036	0.0591	0.2540	0.5585
0.2	0.0025	0.0668	0.2972	0.6068
0.3	0.0027	0.0800	0.3289	0.6352
0.4	0.0038	0.0941	0.3509	0.6522
0.5	0.0051	0.1008	0.3592	0.6580

$\boldsymbol{\theta}$. It was assumed that the missingness of the labels did not depend on the data. Using the likelihood $L_{PC}^{(ig)}(\boldsymbol{\theta})$ that ignores the mechanism for the missing-label indicators, he showed that $\boldsymbol{I}_{PC}(\boldsymbol{\beta})$ can be decomposed as

$$\boldsymbol{I}_{PC}(\boldsymbol{\beta}) = \boldsymbol{I}_{CC}(\boldsymbol{\beta}) - \overline{m}\boldsymbol{I}_{CC}^{(lr)}(\boldsymbol{\beta}), \tag{15}$$

where $\boldsymbol{I}_{CC}(\boldsymbol{\beta})$ is the information about $\boldsymbol{\beta}$ in a completely classified sample $\boldsymbol{x}_{CC}$, $\boldsymbol{I}_{CC}^{(lr)}(\boldsymbol{\beta})$ is the conditional information about $\boldsymbol{\beta}$ under the logistic regression model for the distribution of the class labels given the features in $\boldsymbol{x}_{CC}$, and $\overline{m} = \sum_{j=1}^{n} m_j/n$ is the proportion of unclassified features in the partially classified sample $\boldsymbol{x}_{PC}$. It can be seen from (15) that the loss of information due to the sample being partially classified is equal to $\overline{m}\boldsymbol{I}_{CC}^{(lr)}(\boldsymbol{\beta})$. The consequent decrease in the efficiency in estimating the Bayes' rule can be considerable as illustrated above in Table 1.

In other work on the ARE of $R(\boldsymbol{y}; \hat{\boldsymbol{\beta}}_{PC}^{(ig)})$ compared to $R(\boldsymbol{y}; \hat{\boldsymbol{\beta}}_{CC})$, McLachlan and Gordon (1989) and McLachlan and Scot (1995) evaluated it where the unclassified univariate features had labels that were not MCAR but were *missing (always) at random* (MAR) due to truncation of the features.

In that context, $L_{PC}^{(ig)}(\boldsymbol{\theta})$ is still the appropriate likelihood for the estimation of $\boldsymbol{\theta}$ as the probability that an entity with feature vector $\boldsymbol{y}_j$ has a missing label depends only on $\boldsymbol{y}_j$; that is, the missingness is ignorable. The full likelihood $L_{PC}^{(full)}(\boldsymbol{\Psi})$ actually then reduces to $L_{PC}^{(ig)}(\boldsymbol{\theta})$. Note that the Fisher information is affected by ignoring the missingness in its calculation since the distribution of $L_{PC}^{(ig)}(\boldsymbol{\theta})$ depends on the missingness mechanism (Mealli and Rubin, 2015). However, it can be estimated by the observed information matrix given by the negative of the Hessian of $\log L_{PC}^{(ig)}(\boldsymbol{\theta})$. In the past attention has tended to focus more on when the missingness can be ignored in forming the likelihood rather than forming a full likelihood to exploit the additional information by formulating a model to describe the mechanism underlying the missing data (Harel and Schafer, 2009).

7. Illustration of missing labels with high entropy

In many practical applications cluster labels will be assigned by experts. Manual annotation of the dataset can induce a systematic missingness mechanism. From their examination of some partially classified datasets, Ahfock and McLachlan (2019, 2020) conjectured that in many situations the probability of a particular feature vector being unlabelled is related to the difficulty of classifying the observation correctly. As an example, suppose medical professionals are asked to classify each image from a set of MRI scans into three classes, tumour present, no tumour present, or unknown. It would seem reasonable to expect that the unclassified observations would correspond to images that do not present clear evidence for the presence or absence of a tumour. In such situations the unlabelled images would fall in regions of the feature space where there is class overlap. This led Ahfock and McLachlan (2019, 2020) to argue that in these situations, the unlabelled observations carry additional information that can be used to improve the efficiency of the parameter estimation. They suggested the the difficulty of classifying an observation could be quantified using the Shannon entropy of an entity with feature vector $\boldsymbol{y}$, which is defined by

$$e(\boldsymbol{y}, \boldsymbol{\theta}) = -\sum_{i=1}^{g} \tau_i(\boldsymbol{y}; \boldsymbol{\theta}) \log \tau_i(\boldsymbol{y}; \boldsymbol{\theta}), \tag{16}$$

where $\tau_i(\boldsymbol{y}; \boldsymbol{\theta})$ is the posterior probability (2) that the entity belongs to the ith class given the feature vector $\boldsymbol{y}$. This measure can be used to define regions in the feature space of high class uncertainty.

We present below some results for three datasets that were examined by Ahfock and McLachlan (2019) to demonstrate that the missingness mechanism for the missing labels can be related to the Shannon entropy. For visualisation we find it useful here to work with the negative log entropy.

7.1. Flow cytometry dataset

We first report results for a flow cytometry dataset from Aghaeepour et al. (2013) The dataset consists of fluorescence measurements on $n = 11,792$ cells using $p = 3$ markers. Cluster labels were assigned manually by domain experts using specialised software for the analysis of flow cytometry data. Labels were assigned using a combination of user-defined 'gates' that partition feature space into groups. There were $n_u = 333$ observations that were not assigned to a group at the end of
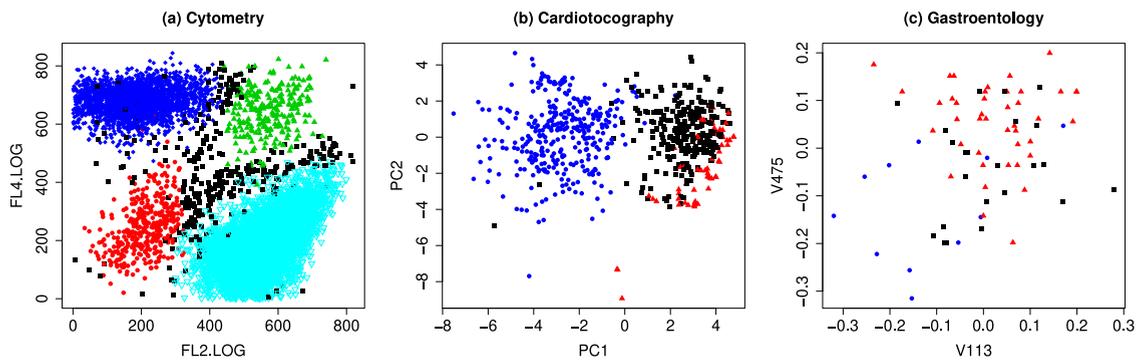
Fig. 1. Partially classified datasets. Black squares denote unlabelled observations. Class membership for labelled observations is represented using different colours and shapes. (a) Flow cytometry dataset. (b) Cardiotocography dataset (c) Gastroentology dataset.

the manual gating process. Fig. 1 shows a bivariate plot of the dataset. Black squares denote unlabelled observations. Clusters are plotted using different colours and shapes. The unlabelled observations appear to be in areas where there is some overlap between clusters. Unlabelled observations appear to be concentrated around class decision boundaries, supporting the general idea that experts will hesitate to label observations that are difficult to classify.

We fitted a skew t-mixture model (Lee and McLachlan, 2014) to estimate the classification entropy of each observation. Row (a) in Fig. 2 compares the negative log entropy of the labelled and unlabelled observations. Panel (i) compares kernel density estimates, and Panel (ii) compares the empirical cumulative distribution functions of the estimated entropy distributions in the labelled and unlabelled groups. Panel (iii) shows a Nadaraya-Watson kernel estimate of the labelling probability. From (i) and (iii), we can see that the unlabelled observations typically have higher entropy than the labelled observations. The estimated missing label probability in (iii) appears to be a decreasing function of the negative log entropy.

7.2. Cardiotocography dataset

We now report results for a second dataset that is a subset of data from Ayres-de Campos et al. (2000) The full dataset consists of 23 features extracted from cardiotocograms on 2126 infants. A panel of three obstetricians used the cardiotocograms to assess fetal state. Observations were labelled as normal, pathological, or suspect given the expert consensus. We take the suspect observations to be unlabelled. The cardiograms were also assigned a morphological pattern $(1, \ldots, 10)$ using automated methods. We restricted attention to the observations with morphological patterns 5,6,9, and 10 as the majority of the unlabelled observations are in these groups. The subset we considered has $n = 670$ observations, with $n_c = 402$ labelled observations and $n_u = 268$ unlabelled observations. We performed dimension reduction using principal components analysis prior to clustering. We worked with the first two principal component scores. Fig. 1 (b) shows the data subset. Normal observations are plotted as blue circles, pathological observations are plotted as red triangles. Suspect observations are plotted as black squares. The bulk of the unlabelled observations are concentrated between the normal and pathological groups. Unlabelled observations appear to be in regions where there is class uncertainty.

We fitted a two-component skew t-mixture model (Lee and McLachlan, 2014) to the observed dataset. We then estimated the entropy of each observation. Row (b) in Fig. 2 compares the negative log entropy of the labelled and unlabelled observations. Panel (i) compares kernel density estimates and panel (ii) compares the empirical cumulative distribution functions. Panel (iii) shows a Nadaraya-Watson kernel estimate of the labelling probability. From (i) and (ii), we can see that the unlabelled observations typically have higher entropy than the labelled observations. The estimated missing label probability in (iii) shows a downward trend with respect to the negative log entropy.

7.3. Gastrointestinal dataset

The third and final dataset to be reported on here concerns a subset of data from Mesejo et al. (2016). The raw dataset consists of 700 features extracted from colonoscopic videos on patients with gastrointestinal lesions. There are $n = 76$ records. A panel of seven doctors reviewed the videos and determined whether the lesions appeared benign or malignant. We formed a consensus labelling using the individual expert labels. The observations where six or more of the experts agreed were treated as labelled giving $n_c = 53$. Observations where fewer than six experts agreed were treated as unlabelled giving $n_u = 23$.

The dataset also includes a ground truth set of labels, obtained using additional histological measurements. The accuracy of the experts can be determined by comparing to the ground truth labels. To reduce the dimension of the dataset, we used sparse linear discriminant analysis (Clemmensen et al., 2011) to select a subset of four features useful for class discrimination using the ground truth labels. These four variables were taken as the features for model based clustering. Fig. 1 (c) shows a bi-variate plot of the data subset. Black squares denote unlabelled observations, red triangles denote benign observations,
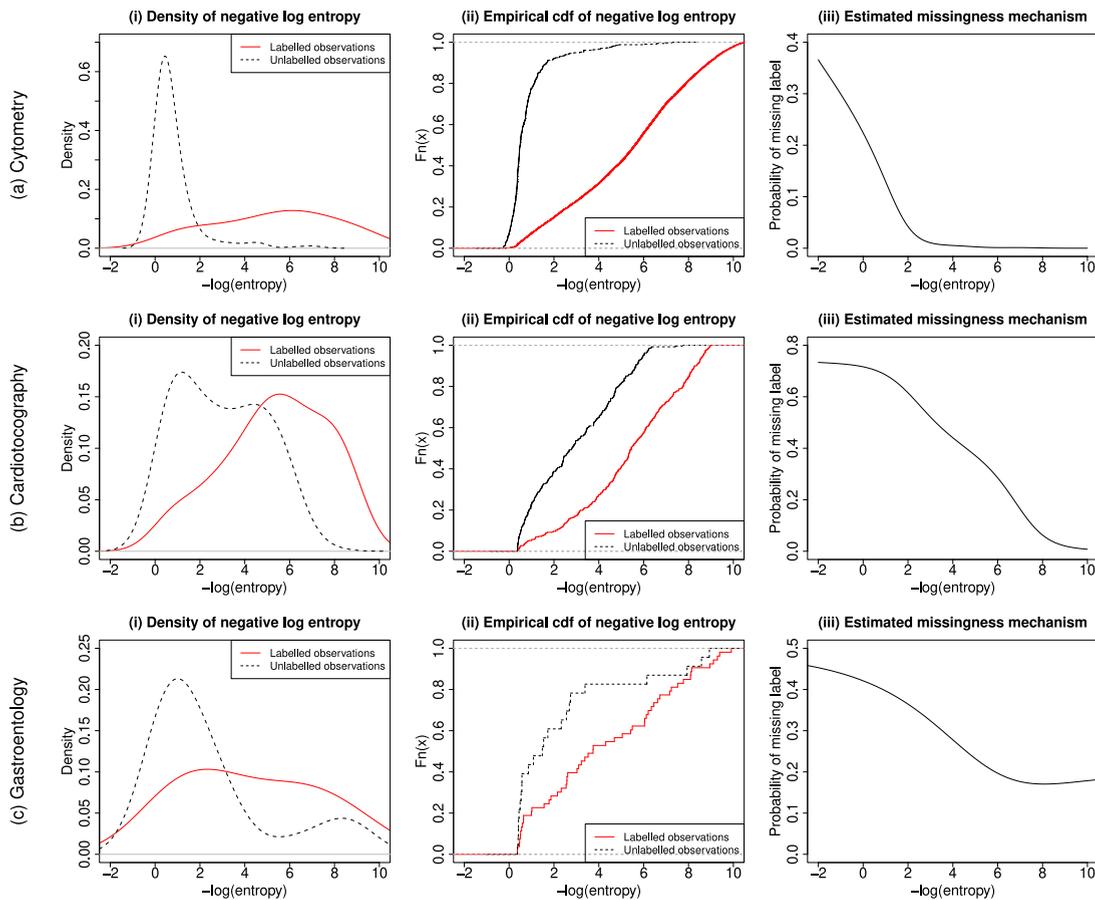
Fig. 2. Analysis of missingness pattern in the partially classified datasets. For each dataset we fitted a model and estimated the entropy of each observation. Row (a) shows results for the cytometry dataset, row (b) shows results for the cardiotocography dataset, and row(c) shows results for the gastroentology dataset. The first column shows density estimates of the negative log entropy. The second column compares the empirical cumulative distribution functions of the negative log entropy. Results for the unlabelled observations are shown as black dashed lines, and results for the labelled observations are shown as red solid lines. The third column shows a Nadaraya-Watson kernel estimate of the labelling probability as a function of the negative log entropy.

and blue circles denote malignant observations. It seems that the unlabelled observations are located in regions where there is group overlap. This dataset is smaller than the cytometry and cardiotocography datasets, so the pattern of missingness is less visually distinctive.

We fitted a two-component t-mixture model to the dataset. We then used the fitted model to estimate the entropy of each observation. Row (c) in Fig. 2 compares the transformed entropy of the labelled and unlabelled observations. Panel (i) compares kernel density estimates and panel (ii) compares the empirical cumulative distribution functions. Panel (iii) shows a Nadaraya-Watson kernel estimate of the labelling probability. From (i) and (ii), we can see that the unlabelled observations typically have higher entropy than the labelled observations. There appears to be a relationship between the entropy and the estimated missing label probability in (iii).

8. Modelling Missingness for Unobserved Class Labels

Following on from the exploratory examination by Ahfock and McLachlan (2019) of partially classified data sets, Ahfock and McLachlan (2020) proposed to treat the labels of the unclassified features as missing data and to introduce a framework for their missingness as in Rubin (1976) for missingness in incomplete-data analysis. Within this framework, they postulated the dependence of the conditional probability that a label is missing given the data by the logistic model with covariate equal to an entropy-based measure. This model was adopted bearing in mind that the unclassified features of many datasets tend to fall in regions of overlap of the classes in the feature space. This is not surprising as entities with features in such regions would tend to be representative of entities that would be difficult to classify correctly, as illustrated for three datasets in the previous section. Ahfock and McLachlan (2020) showed how this dependency on the missingness pattern can be leveraged to provide additional information about the parameters in the optimal classifier specified by the Bayes' rule.
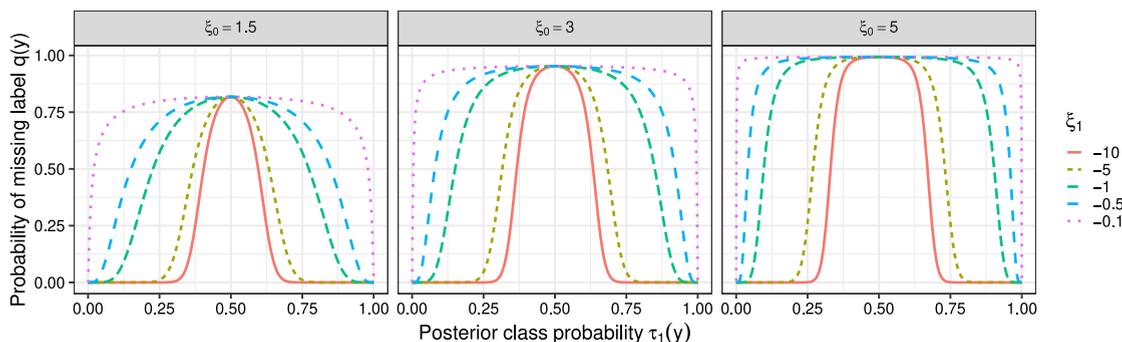
Fig. 3. Probability of a missing label given different parameter values ξ_0, ξ_1 in the proposed missingness model. The probability of a missing label is greatest near the decision boundary where the posterior class probability is 0.5 and decreases as the classification difficulty decreases. The value of ξ_1 controls how rapidly the missingness probability decays moving away from the decision boundary. The value of ξ_0 has an important influence on the maximum probability of missingness that is attained on the decision boundary where the posterior class probability is 0.5.

To this end, they introduced the random variable M_j corresponding to the realized value m_j for the missing-label indicator for the feature vector $\boldsymbol{y}_j$. The missing-data mechanism of Rubin (1976) is specified in the present context by the conditional probabilities

$$\mathrm{pr}\{M_j = m_j \mid \boldsymbol{y}_j, z_j\} = \mathrm{pr}\{M_j = m_j \mid \boldsymbol{y}_j\} \tag{17}$$

$$= q(\boldsymbol{y}_j; \boldsymbol{\Psi}) \qquad (j = 1, \ldots, n), \tag{18}$$

where $\boldsymbol{\Psi} = (\boldsymbol{\theta}^T, \boldsymbol{\xi}^T)^T$ and where the parameter $\boldsymbol{\xi} = (\xi_0, \xi_1)^T$ is distinct from $\boldsymbol{\theta}$.

In the case where the probability (18) does not depend on $\boldsymbol{\theta}$ but only $\boldsymbol{y}_j$ and $\boldsymbol{\xi}$, the missingness is ignorable as discussed in Section 6. But now with this probability depending also on $\boldsymbol{\theta}$, the missingness is non-ignorable.

On adopting the logistic function to model the dependence of the probability (18) on the entropy of the feature vector $\boldsymbol{y}_j$, or more precisely, the negative log entropy, we have that

$$q(\boldsymbol{y}_j; \boldsymbol{\Psi}) = \frac{\exp\{\xi_0 - \xi_1 \log e_j(\boldsymbol{\theta})\}}{1 + \exp\{\xi_0 - \xi_1 \log e_j(\boldsymbol{\theta})\}}, \tag{19}$$

where the parameter $\boldsymbol{\xi} = (\xi_0, \xi_1)^T$ is distinct from $\boldsymbol{\theta}$.

The expected proportion $\gamma(\boldsymbol{\Psi})$ of unclassified features in a partially classified sample $\boldsymbol{x}_{\mathrm{PC}}$ is given by

$$\begin{aligned}
\gamma(\boldsymbol{\Psi}) &= \sum_{j=1}^{n} E(M_j)/n \\
&= E[\mathrm{pr}\{M_j = 1 \mid \boldsymbol{Y}_j\}] \\
&= E\{q(\boldsymbol{Y}; \boldsymbol{\Psi})\}.
\end{aligned} \tag{20}$$

To simplify the numerical computation in the particular case of only $g = 2$ classes with the two-class homoscedastic normal model (3), Ahfock and McLachlan (2020) replaced $\log e_j(\boldsymbol{\theta})$ in (19) by minus the square of the discriminant function $d_j = d(\boldsymbol{y}_j; \boldsymbol{\beta})$ as defined by (4) so that

$$q(\boldsymbol{y}_j; \boldsymbol{\Psi}) = \frac{\exp\{\xi_0 + \xi_1 d(\boldsymbol{y}_j; \boldsymbol{\beta})^2\}}{1 + \exp\{\xi_0 + \xi_1 d(\boldsymbol{y}_j; \boldsymbol{\beta})^2\}}. \tag{21}$$

In support of this approximation, they noted that the negative log entropy is a linear function of d_j^2 ignoring terms of order $O(d_j^{-4})$.

The probability of a missing label can also be expressed in terms of the value of the posterior probability of membership of Class 1, $\tau_1(\boldsymbol{y}_j; \boldsymbol{\beta})$,

$$\mathrm{pr}\{M_j = 1 \mid \boldsymbol{y}_j\} = q(\boldsymbol{y}_j; \boldsymbol{\beta}, \boldsymbol{\xi}) = \frac{\exp\{\xi_0 + \xi_1 [\log(\tau_1(\boldsymbol{y}_j; \boldsymbol{\beta})/\{1 - \tau_1(\boldsymbol{y}_j; \boldsymbol{\beta})\})]^2\}}{1 + \exp\{\xi_0 + \xi_1 [\log(\tau_1(\boldsymbol{y}_j; \boldsymbol{\beta})/\{1 - \tau_1(\boldsymbol{y}_j; \boldsymbol{\beta})\})]^2\}}. \tag{22}$$

Fig. 3 plots the probability of missingness $q(\boldsymbol{y}_j; \boldsymbol{\beta}, \boldsymbol{\xi})$ against the posterior class probability $\tau_1(\boldsymbol{y}_j; \boldsymbol{\beta})$ for the combinations of ξ_0 and ξ_1 that are used in Table 1. The probability of a missing label has no further dependence on $\boldsymbol{\beta}$ given the posterior probability $\tau_1(\boldsymbol{y}_j; \boldsymbol{\beta})$ of membership of Class 1.

9. Full likelihood approach in case of non-ignorable MAR labels

We now consider in the case of non-ignorable missing class labels the estimation of the Bayes' classifier as given by (4) under the two-class homoscedastic normal model (3). We adopt the missing-label mechanism as proposed in Section 8 whereby the probability $q(\boldsymbol{y}_j; \boldsymbol{\Psi})$ of the feature vector $\boldsymbol{y}_j$ having a missing label ($m_j = 1$), is specified by the logistic model (21). Thus the missingness is non-ignorable as this probability now depends also on $\boldsymbol{\theta}$, and hence, the parameters of the Bayes' classifier.

The full likelihood function for $\boldsymbol{\Psi}$ is given by

$$\log L_{\mathrm{PC}}^{(\mathrm{full})}(\boldsymbol{\Psi}) = \log L_{\mathrm{PC}}^{(\mathrm{ig})}(\boldsymbol{\theta}) + \log L_{\mathrm{PC}}^{(\mathrm{miss})}(\boldsymbol{\Psi}), \tag{23}$$

where

$$\log L_{\mathrm{PC}}^{(\mathrm{miss})}(\boldsymbol{\Psi}) = \sum_{j=1}^{n}[(1 - m_j)\log\{1 - q(\boldsymbol{y}_j; \boldsymbol{\Psi})\} + m_j \log q(\boldsymbol{y}_j; \boldsymbol{\Psi})] \tag{24}$$

is the log likelihood function for $\boldsymbol{\Psi}$ formed on the basis of the missing-label indicators m_j ($j = 1, \ldots, n$) and $\log L_{\mathrm{PC}}^{(\mathrm{ig})}(\boldsymbol{\theta})$ is the log likelihood (11) formed ignoring the mechanism for the missing class labels.

We let $\hat{\boldsymbol{\Psi}}_{\mathrm{PC}}^{(\mathrm{full})}$ be the estimate of $\boldsymbol{\Psi}$ formed by consideration of the full likelihood $L_{\mathrm{PC}}^{(\mathrm{full})}(\boldsymbol{\Psi})$ and $\hat{R}_{\mathrm{PC}}^{(\mathrm{full})} = R(\boldsymbol{y}; \hat{\boldsymbol{\beta}}_{\mathrm{PC}}^{(\mathrm{full})})$ be the estimated Bayes' rule obtained by plugging in $\hat{\boldsymbol{\beta}}_{\mathrm{PC}}^{(\mathrm{full})}$ for $\boldsymbol{\beta}$ in the Bayes' rule $R(\boldsymbol{y}; \boldsymbol{\beta})$.

It was noted that there may be an identifiability issue concerning $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ if $\log L_{\mathrm{PC}}^{(\mathrm{miss})}(\boldsymbol{\beta}, \boldsymbol{\xi})$ were to be used on its own for the estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$. But as it is being combined with $\log L_{\mathrm{PC}}^{(\mathrm{ig})}(\boldsymbol{\theta})$ to form the full log likelihood $\log L_{\mathrm{PC}}^{(\mathrm{full})}(\boldsymbol{\Psi})$, $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ are each identifiable with the use of the latter.

The asymptotic relative efficiency (ARE) of the full likelihood rule $\hat{R}_{\mathrm{PC}}^{(\mathrm{full})}$ based on the partially classified sample $\boldsymbol{x}_{\mathrm{PC}}$ compared to the rule $\hat{R}_{\mathrm{CC}}$ based on the completely classified sample $\boldsymbol{x}_{\mathrm{CC}}$ is defined by replacing $\hat{\boldsymbol{\beta}}_{\mathrm{PC}}^{(\mathrm{ig})}$ with $\hat{\boldsymbol{\beta}}_{\mathrm{PC}}^{(\mathrm{full})}$ in (14). In order to derive the ARE of $\hat{R}_{\mathrm{PC}}^{(\mathrm{full})}$, Ahfock and McLachlan (2020) derived the Fisher information about $\boldsymbol{\beta}$ via the full likelihood $L_{\mathrm{PC}}^{(\mathrm{full})}(\boldsymbol{\Psi})$. It is given in the following theorem.

Theorem 1. *The Fisher information about $\boldsymbol{\beta}$ in the partially classified sample $\boldsymbol{x}_{\mathrm{PC}}$ via the full likelihood function $L_{\mathrm{PC}}^{(\mathrm{full})}(\boldsymbol{\Psi})$ can be decomposed as*

$$\boldsymbol{I}_{\mathrm{PC}}^{(\mathrm{full})}(\boldsymbol{\beta}) = \boldsymbol{I}_{\mathrm{CC}}(\boldsymbol{\beta}) - \gamma(\boldsymbol{\Psi})\boldsymbol{I}_{\mathrm{CC}}^{(\mathrm{clr})}(\boldsymbol{\beta}) + \boldsymbol{I}_{\mathrm{PC}}^{(\mathrm{miss})}(\boldsymbol{\beta}), \tag{25}$$

where $\boldsymbol{I}_{\mathrm{CC}}(\boldsymbol{\beta})$ is the information about $\boldsymbol{\beta}$ in the completely classified sample $\boldsymbol{x}_{\mathrm{CC}}$, $\boldsymbol{I}_{\mathrm{CC}}^{(\mathrm{clr})}(\boldsymbol{\beta})$ is the conditional information about $\boldsymbol{\beta}$ under the logistic regression model for the distribution of the class labels given the features in $\boldsymbol{x}_{\mathrm{CC}}$, and $\boldsymbol{I}_{\mathrm{PC}}^{(\mathrm{miss})}(\boldsymbol{\beta})$ is the information about $\boldsymbol{\beta}$ in the missing-label indicators under the assumed logistic model for their distribution given their associated features in the partially classified sample $\boldsymbol{x}_{\mathrm{PC}}$.

It can be seen from (25) that if

$$\boldsymbol{I}_{\mathrm{PC}}^{(\mathrm{miss})}(\boldsymbol{\beta}) > \gamma \boldsymbol{I}_{\mathrm{CC}}^{(\mathrm{clr})}(\boldsymbol{\beta}), \tag{26}$$

then there is actually an increase in the information about $\boldsymbol{\beta}$ in the partially classified sample over the information $\boldsymbol{I}_{\mathrm{CC}}(\boldsymbol{\beta})$ about $\boldsymbol{\beta}$ in the completely classified sample. The inequality in (26) is used in the sense that the left-hand side of the equation, minus the right, is positive definite.

Ahfock and McLachlan (2020) showed that the inequality (26) can hold for various combinations of the parameters, which implies then that $\hat{\boldsymbol{\beta}}_{\mathrm{PC}}$ provides a more efficient estimator of the vector $\boldsymbol{\beta}$ of discriminant function coefficients than $\hat{\boldsymbol{\beta}}_{\mathrm{CC}}$. A consequence of this in such situations is that the asymptotic expected error rate of the estimated Bayes' rule $\hat{R}_{\mathrm{PC}}$ is actually smaller than that of the rule $\hat{R}_{\mathrm{CC}}$ based on a completely classified sample.

The ARE of $\hat{R}_{\mathrm{PC}}^{(\mathrm{full})}$ compared to $\hat{R}_{\mathrm{CC}}$ in the case of equal prior probabilities is given in Theorem 2.

Theorem 2 (Ahfock and McLachlan (2020)). *Under the proposed missing-label model (21), the ARE of $\hat{R}_{\mathrm{PC}}^{(\mathrm{full})}$ compared to $\hat{R}_{\mathrm{CC}}$ is given in the case of $\pi_1 = \pi_2$ by*

$$\mathrm{ARE}(\hat{R}_{\mathrm{PC}}^{(\mathrm{full})}) = 1 - (\Delta^2 + 4)(\gamma d_0 - b_0) \tag{27}$$

for all p, where

$$b_0 = \int_{-\infty}^{\infty} 4\xi_1^2 \Delta^2 y_1^2 q_1(y_1)(1 - q(y_1)) f_{y_1}(y_1) dy_1,$$

$$d_0 = \int_{-\infty}^{\infty} \tau_1(y_1)\tau_2(y_1)q_1(y_1)\gamma^{-1} f_{y_1}(y_1) dy_1,$$

Table 2
Asymptotic relative efficiency of $\hat{R}_{PC}^{(full)}$ for $\pi_1 = \pi_2$ (applicable for all p)

ξ_0	Δ	ξ_1				
		-0.1	-0.5	-1	-5	-10
1.5	1	0.2	1.5	3.6	15.0	23.1
	2	0.8	3.1	4.7	10.3	14.4
	3	1.6	2.9	3.6	6.6	8.9
3	1	0.1	1.0	3.5	20.2	32.5
	2	0.5	4.0	6.4	14.8	20.9
	3	1.9	4.1	5.1	9.4	12.8
5	1	0.01	0.4	2.4	23.4	40.4
	2	0.3	4.4	7.8	19.4	27.5
	3	1.9	5.5	6.9	12.5	16.9

and where

$$\tau_1(y_1) = \text{pr}\{Z = 1 \mid (\boldsymbol{Y})_1 = y_1\} \quad (i = 1, 2),$$

$$q_1(y_1; \Delta, \boldsymbol{\xi}) = \text{pr}\{M = 1 \mid (\boldsymbol{Y})_1 = y_1\},$$

$$f_{y_1}(y_1; \Delta, \pi_1) = \pi_1 \phi(y_1; \Delta/2, 1) + (1 - \pi_1)\phi(y_1; -\Delta/2, 1).$$

In Theorem 2, the term b_0 is a function of $\boldsymbol{I}_{PC}^{(miss)}(\boldsymbol{\beta})$, and the term d_0 is related to the information loss due to the missing-labels $\gamma(\boldsymbol{\Psi})\boldsymbol{I}_{CC}^{(clr)}(\boldsymbol{\beta})$.

In the case of $\pi_1 = \pi_2$, Table 2 gives the ARE of $\hat{R}_{PC}^{(full)}$ compared to $\hat{R}_{CC}$ for various combinations of the parameters Δ, ξ_0, and ξ_1, the results applying for all values of p. It can be seen for most of the combinations in Table 2 that the ARE of $\hat{R}_{PC}^{(full)}$ is greater than one, being appreciably greater than one for some combinations of the parameters. For example, for $\Delta = 1$ (representing classes close together) or $\Delta = 2$ (classes moderately separated), the ARE is not less than 14.8 for any combination with $\xi_0 = 3$ or 5 and $\xi_1 = -5$ or -10, being as high as 40.4 for $\Delta = 1, \xi_0 = 5, \xi_1 = -10$. This shows that the asymptotic expected excess error rate using the partially classified sample $\boldsymbol{x}_{PC}$ can be much lower than the corresponding excess rate using the completely classified sample $\boldsymbol{x}_{CC}$.

The general expression for the ARE of $\hat{R}_{PC}^{(full)}$ for $\pi_1 \neq \pi_2$ is available in the supplementary material of Ahfock and McLachlan (2020). They noted that this ARE is not sensitive to the value π_1 in the range (0.2, 0.8), so that Theorem 2 can provide useful guidelines for arbitrary prior probabilities.

10. ARE of $\hat{R}_{PC}^{(ig)}$ in the case of non-ignorable MAR labels

Under the proposed missingness model (21), the labels are missing at random but the distinctness assumption (Little and Rubin, 2014) does not hold as the parameter vector $\boldsymbol{\theta}$ of the generative model enters the likelihood for the missing-label indicators. Even though the missingness mechanism is non-ignorable, the use of $L^{(ig)}(\boldsymbol{\theta})$ still yields a consistent estimator, although it is no longer fully efficient. As $L^{(ig)}(\boldsymbol{\theta})$ is a partial likelihood function, the asymptotic covariance matrix of the resulting estimator can still be determined by taking the inverse of the expectation of the negative of the Hessian matrix (Little and Rubin, 2014; Little et al., 2017).

In Lemma 1, we give an expression for $\boldsymbol{I}_{PC}^{(ig)}(\boldsymbol{\beta})$ obtained on noting that

$$\boldsymbol{I}_{PC}^{(ig)}(\boldsymbol{\beta}) = \boldsymbol{I}_{PC}^{(full)}(\boldsymbol{\beta}) - \boldsymbol{I}_{PC}^{(miss)}(\boldsymbol{\beta}).$$

and using (25) in Theorem 1.

Lemma 1. *The information about $\boldsymbol{\beta}$ in the partially classified sample $\boldsymbol{x}_{PC}$ using the likelihood that ignores the missingness mechanism $L_{PC}^{(ig)}(\boldsymbol{\Psi})$ can be decomposed as*

$$\boldsymbol{I}_{PC}^{(ig)}(\boldsymbol{\beta}) = \boldsymbol{I}_{CC}(\boldsymbol{\beta}) - \gamma(\boldsymbol{\Psi})\boldsymbol{I}_{CC}^{(clr)}(\boldsymbol{\beta}), \tag{28}$$

where $\boldsymbol{I}_{CC}(\boldsymbol{\beta})$ is the information about $\boldsymbol{\beta}$ in the completely classified sample $\boldsymbol{x}_{CC}$ and where $\boldsymbol{I}_{CC}^{(clr)}(\boldsymbol{\beta})$ is the conditional information about $\boldsymbol{\beta}$ under the logistic regression model for the distribution of the class labels given the features in $\boldsymbol{x}_{CC}$.

The ARE of $\hat{R}_{PC}^{(ig)}$ compared to $\hat{R}_{CC}$ in the case of equal prior probabilities is given in Theorem 3.

Theorem 3. *Under the proposed missing-label model, the ARE of $\hat{R}_{PC}^{(ig)}$ compared to $\hat{R}_{CC}$ is given in the case of $\pi_1 = \pi_2$ by*

$$\text{ARE}(\hat{R}_{PC}^{(ig)}) = 1 - (\Delta^2 + 4)\gamma d_0 \tag{29}$$

Table 3
Asymptotic relative efficiency of $\hat{R}_{PC}^{(ig)}$ compared to $\hat{R}_{PC}^{(full)}$ for $\pi_1 = \pi_2$ (applicable for all p)

ξ_0	Δ	ξ_1				
		-0.1	-0.5	-1	-5	-10
1.5	1	0.81	0.18	0.09	0.04	0.03
	2	0.39	0.14	0.12	0.07	0.06
	3	0.32	0.23	0.20	0.13	0.10
3	1	0.78	0.09	0.04	0.02	0.02
	2	0.29	0.07	0.06	0.04	0.04
	3	0.21	0.13	0.12	0.08	0.07
5	1	0.83	0.05	0.02	0.01	0.01
	2	0.41	0.04	0.03	0.03	0.02
	3	0.19	0.08	0.08	0.06	0.05

for all p, where d_0 is defined as in Theorem 2.

The result follows from using Lemma 1 and Theorem 1. The term $b_0 = 0$ in Theorem 2 is now zero since $\boldsymbol{I}_{PC}^{(miss)}(\boldsymbol{\beta})$ does not contribute to $\boldsymbol{I}_{PC}^{(ig)}(\boldsymbol{\beta})$.

The asymptotic relative efficiency of $\hat{R}_{PC}^{(ig)}$ compared to $\hat{R}_{PC}^{(full)}$ can now be obtained by combining the results in Theorems 2 and 3. It is given in Theorem 4 in the case of equal prior probabilities.

Theorem 4. *Under the proposed missing-label model, the ARE of $\hat{R}_{PC}^{(ig)}$ compared to $\hat{R}_{PC}^{(full)}$ is given in the case of $\pi_1 = \pi_2$ by*

$$\text{ARE}(\hat{R}_{PC}^{(ig)} : \hat{R}_{PC}^{(full)}) = \frac{\text{ARE}(\hat{R}_{PC}^{(ig)})}{\text{ARE}(\hat{R}_{PC}^{(full)})}, \tag{30}$$

where the numerator $\text{ARE}(\hat{R}_{PC}^{(ig)})$ is given by Theorem 3 and the denominator $\text{ARE}(\hat{R}_{PC}^{(full)})$ is given by Theorem 2.

Table 3 reports the ARE of $\hat{R}_{PC}^{(ig)}$ for different combinations of Δ, ξ_0, and ξ_1. This ARE is very low for most of the combinations, supporting the idea that incorporating the missingness mechanism can supply additional information and improve classification accuracy. Indeed, except for those combinations of the parameters with $\Delta < 3$ and $\xi_1 = -0.1$ in Table 3, the ARE is very low. It can be seen for fixed values of Δ and ξ_0, the ARE decreases with ξ_1. Thus the case $\xi_1 = -0.1$ corresponds to a situation where the probability of a feature vector $\boldsymbol{y}_j$ having a missing label depends only weakly on the absolute value of the discriminant function at this value of $\boldsymbol{y}_j$.

10.1. Simulation

We conducted a simulation to assess to what extent the asymptotic results of the previous section apply in practice. For each of the combinations of the parameters in Table 1, we generated $B = 1000$ samples of $\boldsymbol{X} = (\boldsymbol{Y}^T, Z)^T$ to form the completely classified sample $\boldsymbol{x}_{CC}$ and the partially classified sample $\boldsymbol{x}_{PC}$. On each replication, the estimates $\hat{\boldsymbol{\beta}}_{PC}^{(ig)}$ and $\hat{\boldsymbol{\beta}}_{PC}^{(full)}$ were computed using a quasi-Newton algorithm, along with the conditional error rates, $\text{err}(\hat{\boldsymbol{\beta}}_{PC}^{(ig)}; \boldsymbol{\beta})$ and $\text{err}(\hat{\boldsymbol{\beta}}_{PC}^{(full)}; \boldsymbol{\beta})$. We let $\text{err}(\hat{\boldsymbol{\beta}}_{PC}^{(ig,b)}; \boldsymbol{\beta})$ and $\text{err}(\hat{\boldsymbol{\beta}}_{PC}^{(full,b)}; \boldsymbol{\beta})$ denote the conditional error rate of $\hat{R}_{PC}^{(ig)}$ and of $\hat{R}_{PC}^{(full)}$, respectively, on the bth replication.

The relative efficiency (RE) of $\hat{R}_{PC}^{(ig)}$ compared to $\hat{R}_{PC}^{(full)}$ was estimated by

$$\overline{\text{RE}}(\hat{R}_{PC}^{(ig)} : \hat{R}_{PC}^{(full)}; \boldsymbol{\beta}) = \frac{B^{-1}\sum_{b=1}^{B}\{\text{err}(\hat{\boldsymbol{\beta}}_{PC}^{(full,b)}; \boldsymbol{\beta}) - \text{err}(\boldsymbol{\beta})\}}{B^{-1}\sum_{b=1}^{B}\{\text{err}(\hat{\boldsymbol{\beta}}_{PC}^{(ig,b)}; \boldsymbol{\beta}) - \text{err}(\boldsymbol{\beta})\}}. \tag{31}$$

Table 4 reports the results for $n = 500$, and Table 5 reports the results for $n = 100$. As expected, the agreement with the theoretical asymptotic results is better for $n = 500$ than with $n = 100$. For $n = 100$, the simulated relative efficiency is greater than one for the simulations with $\Delta = 1, \xi_1 = -0.1, \xi_0 = 1.5$ and $\Delta = 1, \xi_1 = -0.1, \xi_0 = 5$. The agreement between the simulation results and the theoretical values is good for $n = 500$ and $\xi_1 \geq 1$.

11. Discussion

The results presented here on a statistical SSL approach to classification are quite encouraging in that they show there are gains to be made in efficiency concerning the estimation of the parameters in the Bayes' rule. Situations where these gains can be achieved are where the labels of the unclassified features in the partially classified sample are not missing completely at random (MCAR) but rather occur according to a missing-label mechanism that is able to be modelled appropriately.

Table 4

Simulated relative efficiency of $\hat{R}_{PC}^{(ig)}$ compared to $\hat{R}_{PC}^{(full)}$ with $\pi_1 = \pi_2$ for $n = 500$, $p = 1$ (Bootstrap standard errors are in parentheses)

ξ_0	Δ	ξ_1				
		-0.1	-0.5	-1	-5	-10
1.5	1	0.93 (0.026)	0.18 (0.011)	0.09 (0.005)	0.04 (0.002)	0.03 (0.002)
	2	0.39 (0.021)	0.16 (0.009)	0.12 (0.007)	0.08 (0.005)	0.06 (0.004)
	3	0.34 (0.018)	0.22 (0.012)	0.20 (0.011)	0.14 (0.009)	0.12 (0.008)
3	1	0.98 (0.041)	0.09 (0.005)	0.04 (0.002)	0.02 (0.001)	0.02 (0.001)
	2	0.31 (0.018)	0.08 (0.005)	0.06 (0.003)	0.04 (0.003)	0.03 (0.002)
	3	0.21 (0.011)	0.13 (0.008)	0.12 (0.008)	0.09 (0.006)	0.08 (0.005)
5	1	0.88 (0.028)	0.14 (0.015)	0.02 (0.001)	0.01 (0.001)	0.01 (0.001)
	2	0.44 (0.032)	0.03 (0.002)	0.03 (0.002)	0.03 (0.002)	0.02 (0.002)
	3	0.21 (0.013)	0.08 (0.005)	0.08 (0.005)	0.07 (0.004)	0.06 (0.004)

Table 5

Simulated relative efficiency of $\hat{R}_{PC}^{(ig)}$ compared to $\hat{R}_{PC}^{(full)}$ with $\pi_1 = \pi_2$ for $n = 100$, $p = 1$ (Bootstrap standard errors are in parentheses)

ξ_0	Δ	ξ_1				
		-0.1	-0.5	-1	-5	-10
1.5	1	1.12 (0.037)	0.27 (0.029)	0.10 (0.006)	0.04 (0.003)	0.03 (0.002)
	2	0.42 (0.020)	0.16 (0.009)	0.13 (0.008)	0.10 (0.007)	0.07 (0.004)
	3	0.33 (0.018)	0.26 (0.015)	0.25 (0.014)	0.25 (0.015)	0.29 (0.016)
3	1	1.05 (0.018)	0.42 (0.028)	0.09 (0.015)	0.02 (0.002)	0.02 (0.001)
	2	0.43 (0.046)	0.07 (0.005)	0.06 (0.003)	0.06 (0.004)	0.05 (0.003)
	3	0.24 (0.015)	0.16 (0.009)	0.15 (0.009)	0.17 (0.009)	0.20 (0.012)
5	1	0.96 (0.019)	0.79 (0.033)	0.28 (0.024)	0.01 (0.001)	0.01 (0.001)
	2	0.92 (0.073)	0.04 (0.003)	0.04 (0.002)	0.04 (0.003)	0.04 (0.003)
	3	0.19 (0.012)	0.10 (0.006)	0.10 (0.007)	0.15 (0.009)	0.19 (0.012)

A common situation with labels that are missing at random (MAR) but non-ignorable is one in which the features that are difficult to classify correctly are left unclassified. In the theoretical and practical results presented in this review the missingness mechanism has been represented via the use of the logistic function to model the conditional probability that a label is missing via its dependence on the log entropy of the feature vector. The choice of the logistic function can be modified to handle situations where the features with missing labels are confined to lying in only certain regions of high entropy.

As remarked in the introduction, the fact that the full likelihood-based rule $\hat{R}_{PC}^{(full)}$ can outperform the completely super-vised learning (CSL) rule $\hat{R}_{CC}$ is very surprising. In situations where this can occur, that is, where the unclassified features are those occurring in regions of high entropy in the feature space, means that a large number of unclassified observations are not needed to build a SSL classifier with small error rate.

Related to the full likelihood SSL classifier having performance better or comparable to the CSL classifier is that the SSL rule $\hat{R}_{PC}^{(ig)}$ that ignores the missingness mechanism for the missing labels performs very poorly relative to the full likelihood-based rule $\hat{R}_{PC}^{(full)}$. The asymptotic relative efficiency of $\hat{R}_{PC}^{(ig)}$ compared to $\hat{R}_{PC}^{(full)}$ has been derived here using the results produced in Ahfock and McLachlan (2020) during their derivation of the ARE of $\hat{R}_{PC}^{(full)}$. The results so obtained for the ARE of $\hat{R}_{PC}^{(ig)}$ show how it can be in situations where the missing-label pattern is non-ignorable.

Simulation results undertaken by Lachenbruch (1966, 1974) and theoretical results provided initially by McLachlan (1972) and, more recently, Cannings et al. (2020), have shown that the error rate of a classifier can be adversely affected by misclassification of its training data. Thus in situations where the SSL rule based on the full likelihood $\hat{R}_{PC}^{(full)}$ has better or comparable performance to the completely classified (CSL) rule $\hat{R}_{CC}$, one would expect it to do even better for a partially classified sample in which some of the labelled features are misclassified.

In the results presented in this paper, the error rate of a classifier refers to its application to unclassified features not in the partially classified sample. But in ongoing work, we also have been considering the error rates of the classifiers in their application to the unclassified features in the partially classified training sample. Preliminary results show that the relative efficiency of the rules $\hat{R}_{PC}^{(ig)}$ and $\hat{R}_{PC}^{(full)}$ has not significantly changed. In the situations where the unclassified features in the sample are those that are the most difficult to classify, it is to be expected the actual sizes of the error rates of these rules are greater than in their application to new randomly chosen features. Also, the rule $\hat{R}_{PC}^{(full)}$ will not be superior to the CSL rule $\hat{R}_{CC}$ since the latter is being applied to the unclassified features whose true labels were used in the construction of this CSL rule. This provides the rule $\hat{R}_{CC}$ with too great an advantage over the rule $\hat{R}_{PC}^{(full)}$.

Attention has been confined here to the situation where the features are from a mixture of two classes that have mul-tivariate normal distributions with a common covariance matrix. In this situation, it is possible to investigate analytically

the relative performance of the classifiers. However, as the model of Ahfock and McLachlan (2020) for the missingness of the class labels is defined in terms of the entropy of an observed feature, it is applicable to more than two classes with arbitrary distributions. As part of our ongoing research on the full likelihood approach of Ahfock and McLachlan (2020), we intend to carry out simulations in such situations, in particular, for heteroscedastic multivariate normal distributions.

Acknowledgement

References

Aghaeepour, N., Finak, G., FLOW-Consortium, DREAM-Consortium, Hoos, H., Mosmann, T.R., Brinkman, R., Gottardo, R., Scheuermann, R.H., 2013. Critical assessment of automated flow cytometry data analysis techniques. Nature Methods 10, 228–238.

Ahfock, D., McLachlan, G.J., 2019. On missing data patterns in semi-supervised learning. arXiv e-prints. arXiv:1904.02883.

Ahfock, D., McLachlan, G.J., 2020. An apparent paradox: A classifier trained from a partially classified sample may have smaller expected error rate than that if the sample were completely classified. Statistics and Computing 30, 1779–1790.

Belkin, M., Niyogi, P., Sindhwani, V., 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research 7, 2399–2434.

Berthelot, D., Carlini, N., Cubuk, E.D., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C., Sohn, K., 2020. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. arXiv e-prints. arXiv:1911.09785.

Berthelot, D., Carlini, N., Goodfellow, I., 2019. Mixmatch: A holistic approach to semi-supervised learning. In: Advances in neural information processing systems 32.

Ayres-de Campos, D., Bernardes, J., Garrido, A., Marques-de S, J., Pereira-Leite, L., 2000. Sisporto 2.0: A program for automated analysis of cardiotocograms. The Journal of Maternal Fetal Medicine 9, 311–318.

Castelli, V., Cover, T.M., 1996. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. IEEE Transactions on Information Theory 42, 2102–2117.

Chapelle, O., Schlökopf, B., Zien, A., 2010. Semi-Supervised Learning. MIT Press, Cambridge, MA.

Chen, S., Dobriban, E., Lee, J., 2020a. A group-theoretic framework for data augmentation. In: Advances in neural information processing systems 33.

Chen, Z., Dutton, B., Ramachandra, B., Wu, T., Vatsavai, R.R., 2020b. Local clustering with mean teacher for semi-supervised learning. arXiv e-prints. arXiv:2004.09665.

Clemmensen, L., Hastie, T., Witten, D., Ersbll, B., 2011. Sparse discriminant analysis. Technometrics 406–413.

Decoste, D., Schökopf, B., 2002. Training invariant support vector machines. Machine Learning 46, 161–190.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society B 39, 1–38.

Efron, B., 1975. The efficiency of logistic regression compared to normal discriminant analysis. Journal of the American Statistical Association 70, 892–898.

Gallaugher, M., McNicholas, P.D., 2019. On fractionally-supervised classification: weight selection and extension to the multivariate t-distribution. Journal of Classification 36, 232–265.

Ganesalingam, S., McLachlan, G.J., 1978. The efficiency of a linear discriminant function based on unclassified initial samples. Biometrika 65, 658–665.

Göpfert, C., Ben-David, S., Bousquet, O., Gelly, S., Tolstikhin, I., Urner, R., 2019. When can unlabeled data improve the learning rate? In: Conference on learning theory, pp. 1500–1518.

Grandvalet, Y., Bengio, Y., 2005. Semi-supervised learning by entropy minimization. In: Advances in neural information processing systems 17, pp. 529–536.

Harel, O., Schafer, J., 2009. Partial and latent ignorability in missing-data problems. Biometrika 96, 37–50.

Hartley, H.O., Rao, J.N.K., 1968. Classification and estimation in analysis of variance problems. International Statistical Review 36, 141–147.

Hills, M., 1966. Allocation rules and their error rates (with discussion). Journal of the Royal Statistical Society B 28, 1–31.

Horlava, N., Mironenko, A., Niehaus, S., Wagner, S., Roeder, I., Scherf, N., 2020. A comparative study of semi- and self-supervised semantic segmentation of biomedical microscopy data. arXiv e-prints. arXiv:2011.08076.

Ji, D., Smyth, P., Steyvers, M., 2020. Can i trust my fairness metric? Assessing fairness with unlabeled data and bayesian inference. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (Eds.), Advances in neural information processing systems 33.

Lachenbruch, P.A., 1966. Discriminant analysis when the initial samples are misclassified. Technometrics 8, 657–662.

Lachenbruch, P.A., 1974. Discriminant analysis when the initial samples are misclassified II: Non-random misclassification models. Technometrics 16, 419–424.

Lafferty, J., Wasserman, L., 2007. Statistical analysis of semi-supervised regression. In: Advances in neural information processing systems 20, pp. 801–808.

Laine, S., Aila, T., 2017. Temporal ensembling for semi-supervised learning. In: 5th international conference on learning representations.

Lee, S., McLachlan, G.J., 2014. Finite mixtures of multivariate skew t-distributions: Some recent and new results. Statistics and Computing 24, 181–202.

McLachlan, G.J., 1972. Asymptotic results for discriminant analysis when the initial samples are misclassified. Technometrics 14, 415–422.

McLachlan, G.J., 1975. Iterative reclassification procedure for constructing and asymptotically optimal rule of allocation in discriminant analysis. Journal of the American Statistical Association 70, 365–369.

McLachlan, G.J., 1977. Estimating the linear discriminant function from initial samples containing a small number of unclassified observations. Journal of the American Statistical Association 72, 403–406.

McLachlan, G.J., 1982. The classification and mixture maximum likelihood approaches to cluster analysis. In: Handbook of statistics, vol. 2. North-Holland, Amsterdam, pp. 199–208.

McLachlan, G.J., 1992. Discriminant Analysis and Statistical Pattern Recognition. Wiley, New York.

McLachlan, G.J., Ahfock, D., 2021. Estimation of classification rules from partially classified data. In: Chadjipadelis, T., Lausen, B., Markos, A., Lee, T.R., Montanari, A., Nugent, R. (Eds.), Data analysis and rationality in a complex world, studies in classification, data analysis, and knowledge organization. Springer, Berlin, pp. 149–157.

McLachlan, G.J., Basford, K.E., 1988. Mixture Models: Inference and Applications to Clustering. Marcel Dekker, New York.

McLachlan, G.J., Gordon, R.D., 1989. Mixture models for partially unclassified data: A case study of renal venous renin levels in essential hypertension. Statistics in Medicine 8, 1291–1300.

McLachlan, G.J., Krishnan, T., 2008. The EM algorithm and Extensions, 2nd Wiley, Hoboken, New Jersey.

McLachlan, G.J., Scot, D., 1995. On the asymptotic relative efficiency of the linear discriminant function under partial nonrandom classification of the training data. Journal of Statistical Computation and Simulation 52, 415–442.

Mealli, F., Rubin, D.B., 2015. Clarifying missing at random and related definitions, and implications when coupled with exchangeability. Biometrika 102, 995–1000.

Mesejo, P., Pizarro, D., Abergel, A., Rouquette, O., Beorchia, S., Poincloux, L., Bartoli, A., 2016. Computer-aided classification of gastrointestinal lesions in regular colonoscopy. IEEE Transactions on Medical Imaging 35, 2051–2063.

Miyato, T., Maeda, S., Koyama, M., Ishii, S., 2019. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 41, 1979–1993.

Nair, V., Alonso, J.F., Beltramelli, T., 2019. Realmix: Towards realistic semi-supervised deep learning algorithms. arXiv e-prints. arXiv:1912.08766.

Ng, A.Y., Jordan, M.I., 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In: Advances in neural information processing systems 14, pp. 841–848.

O'Neill, T.J., 1978. Normal discrimination with unclassified observations. Journal of the American Statistical Association 70, 365–369.

Rubin, D.B., 1976. Inference and missing data. Biometrika 63, 581–592.

Singh, A., Nowak, R.D., Zhu, X., 2008. Unlabeled data: Now it helps, now it doesn't. In: Advances in neural information processing systems 21, pp. 1513–1520.

Smith, C.A.B., 1966. Contribution to the discussion of paper by M. Hills. Journal of the Royal Statistical Society B 28, 21.

Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E., Kurakin, A., Zhang, H., Raffel, C., 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. arXiv e-prints. arXiv:2001.07685.

Sokolovska, N., Cappe, O., Yvon, F., 2008. The asymptotics of semi-supervised learning in discriminative probabilistic models. In: Proceedings of the 25th international conference on machine learning, pp. 984–991.

Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in neural information processing systems 30, pp. 1195–1204.

Teo, C.H., Globerson, A., Roweis, S.T., Smola, A.J., 2008. Convex learning with invariances. In: Advances in neural information processing systems 20, pp. 1489–1496.

Vbrik, I., McNicholas, P.D., 2015. Fractionally-supervised classification. Journal of Classification 32, 359–381.

Wang, Y., Guo, J., Song, S., Huang, G., 2020. Meta-semi: A meta-learning approach for semisupervised learning. arXiv e-prints. arXiv:2007.02394.

Wei, X., Wei, X., Kong, X., Lu, S., Xing, W., Lu, W., 2021. Fmixcutmatch for semi-supervised deep learning. Neural Networks 133, 166–176.

Xie, Q., Dai, Z., Hovy, E., Luong, M.-T., Le, Q.V., 2019. Unsupervised data augmentation for consistency training. arXiv e-prints. arXiv:1904.12848.

Yuval, O., Rosset, S., 2020. Semi-supervised empirical risk minimization: When can unlabeled data improve prediction. arXiv e-prints. arXiv:2009.00606.